Proyecto 2 – BI

Juan Esteban Vergara
Jessica A. Robles Moreno
María Paula González

Tabla de contenido

Comprensión de negocio y necesidades analíticas:

Para este proyecto se trabajó con la iniciativa InfraestructuraVisible de la Universidad de los Andes que da libre acceso a la información de la infraestructura colombiana y sus indicadores. El análisis gira entorno al impacto que tiene la minería en los indicadores socioeconómicos de Colombia. Revisando los comentarios de los proyectos de semestres pasados y las entrevistas con el cliente se vio que se mencionaba recurrentemente el impacto en población, educación y violencia. Con esto en mente se buscó relación entre estos indicadores, la minería y la población a través de investigaciones, textos de análisis de impacto social y noticias relacionadas tanto en Colombia como a nivel Latinoamérica. Todas las fuentes mencionadas se buscaron en países con contextos sociales similares a Colombia para no basarnos en criterios descontextualizados debido a factores como cultura y región.

Atributos dimensión educación

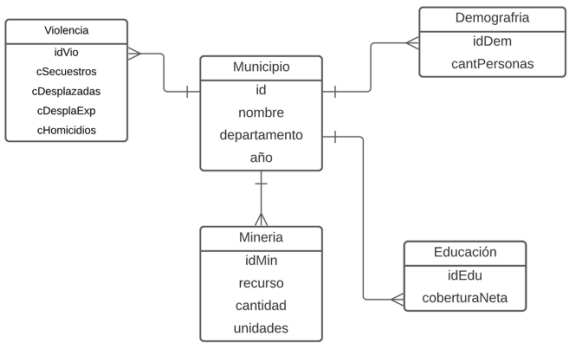
- idEdu -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.

- coberturaNeta -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cobertura.

Atributos dimensión minería

- idMin -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.
- recurso -> Tipo 3: Casi no se usa para realizar cambios debido a la naturaleza del negocio de mineria.
- Cantidad -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad producida.
- Unidades -> Tipo 0: Nunca se cambia, las unidades de medida se mantienen constantes. No hay alteraciones en este atributo.

Tabla de hecho Municipio para el escenario de violencia:



Granularidad: Los datos de la tabla de hecho tienen baja granularidad debido a que tienen información puntual como año, departamento y nombre. Con los datos obtenidos del negocio se encontró que se podía aumentar la granularidad de la fecha con la información de mes y trimestre, pero se observó que era inconsistente y además datos como los de violencia solo contaban con el año por lo que no se aumentó la granularidad para poder realizar comparaciones de forma más ágil. La llave foránea está compuesta por las 3 medidas seleccionadas por lo que un nivel de granularidad es mayor.

Medidas:
Las 3 medidas “Nombre”, “Departamento” y “Año” son semi-aditivas ya que no son operables como dato numérico, pero son medidas agrupables y operables por cantidad de repeticiones.
El “id” es no-aditivo ya que no se realiza ninguna operación sobre este.

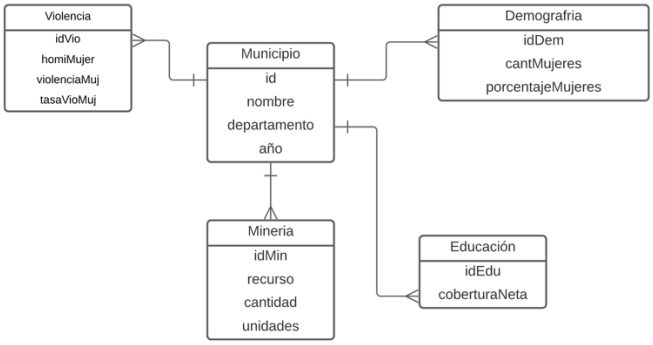
Atributos dimensión demografía

- idDem -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.
- cantPersonas -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de personas.

Atributos dimensión violencia

- idVio -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.
- cSecuestros -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de secuestros.
- cDespazadas -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de personas desplazadas.
- cDesplaExp -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de personas desplazadas expulsadas.
- cHomicidios -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de homicidios.

Tabla de hecho Municipio para el escenario de mujeres:



Granularidad: Los datos de la tabla de hecho tienen baja granularidad debido a que tienen información puntual como año, departamento y nombre. Con los datos obtenidos del negocio se encontró que se podía aumentar la granularidad de la fecha con la información de mes y trimestre, pero se observó que era inconsistente y además datos como los de violencia solo contaban con el año por lo que no se aumentó la granularidad para poder realizar comparaciones de forma más ágil. La llave foránea está compuesta por las 3 medidas seleccionadas por lo que un nivel de granularidad es mayor.

Medidas:
Las 3 medidas “Nombre”, “Departamento” y “Año” son semi-aditivas ya que no son operables como dato numérico, pero son medidas agrupables y operables por cantidad de repeticiones.
El “id” es no-aditivo ya que no se realiza ninguna operación sobre este.

Atributos dimensión demografía

- idDem -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.
- cantMujeres -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de cantidad de mujeres.
- porcentajeMujeres -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico del porcentaje de mujeres.

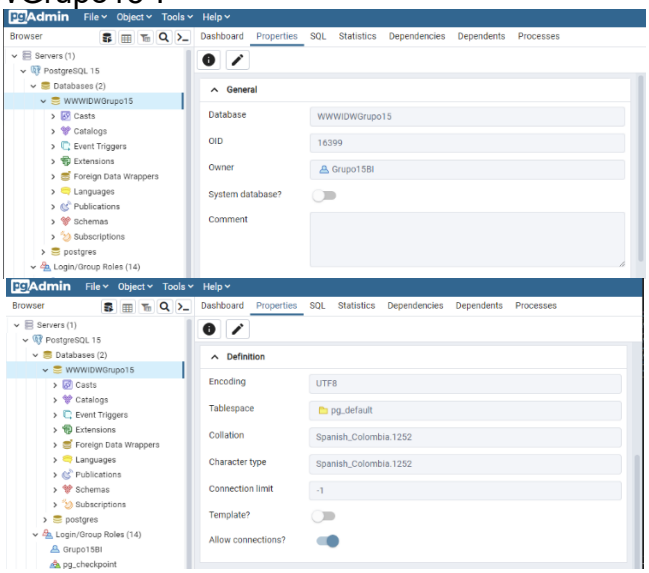
Atributos dimensión violencia

- idVio -> Tipo 0: nunca se cambia, ya que el id representa el registro puntual. No representa sus alteraciones.
- homiMujer -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de los valores de homicidio hacia mujeres.
- violenciaMuj -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de los valores de violencia hacia mujeres.
- tasVioMuj -> Tipo 4: Se crea una dimensión para los nuevos, ya que se almacena el histórico de la tasa de violencia hacia mujeres.

Descripción del proceso

1. Creación de la base de datos

La base de datos que implementamos fue la que se creó para el Laboratorio 5, llamada “WWWIDWGrupo15”.



2. Ejecutar Airflow entender su funcionamiento

Para poder ejecutar Airflow fue necesario comenzar la ejecución de Docker Desktop, al iniciar esta aplicación comienza la ejecución de Airflow mediante un contenedor de Docker, y para ir a la interfaz fue necesario dirigirnos a la dirección http://localhost:8080/. Para encontrar los archivos con los que trabaja esta aplicación fue necesario colocar los archivos csv anteriormente creados en el jupyter.

3. Perfilamiento de datos

Se revisaron los datos proporcionados por la empresa para el laboratorio. Entre estos nos podemos encontrar con once archivos Excel y un archivo json. Para poder revisar mejor su contenido y hacer una mejor limpieza de estos decidimos analizarlos desde dataframes, y así poder utilizar funciones de análisis que ofrece la librería pandas y poder mantener un registro de los cambios hechos y las visualizaciones de estos.

Para el archivo minerias.json encontramos que tiene 45215 registros y 11 columnas. Este archivo no tenía valores nulos y es el eje central del proyecto. Es por esto que lo utilizamos para hacer la limpieza de los otros datos uniendo sus valores por el municipio, departamento y año.

Para el archivo Conficto Armado.xlsx encontramos que tiene 314119 registros y 13 columnas. Aquí se encuentra la información de la violencia de la mujer, secuestros y homicidios, es por esto que fue importante para nuestro uso. Para hacer la limpieza de este lo que se hizo fue buscar los valores de los indicadores que nos interesaban y omitimos los registros nulos, ya que asumimos que no existían registros valiosos.

Para el archivo Demografía y Poblacion.xlsx encontramos que tiene 897998 registros y 13 columnas. Aquí se encentrábamos la cantidad de personas y la cantidad de mujeres en los municipios. Para hacer la limpieza de este lo que se hizo fue buscar los valores de los indicadores que nos interesaban y omitimos los registros nulos, ya que asumimos que no existían registros valiosos.

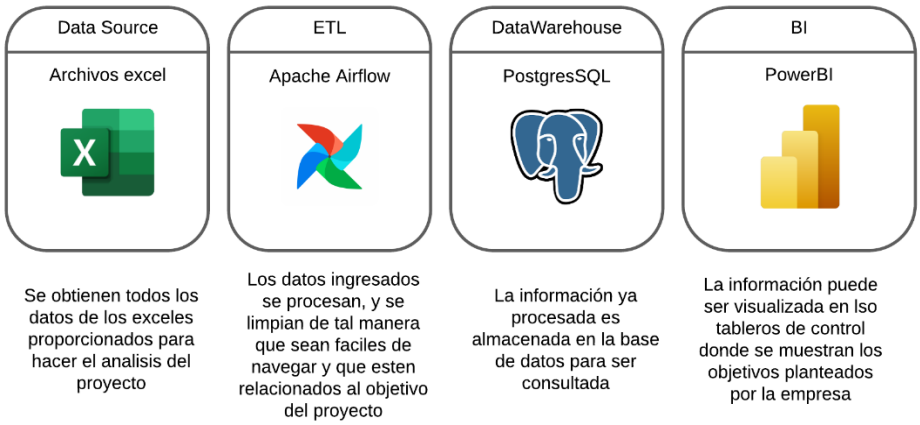
Para el archivo Educacion.xlsx encontramos que tiene 291957 registros y 13 columnas.

A los siguientes archivos no se les hizo una limpieza de datos ya que consideramos que los datos suministrados en ellos no eran de gran relevancia para los modelos dimensionales que se querían implementar.

Para el archivo Medicion de Desempleo Departamental.xlsx encontramos que tiene 100 registros y 13 columnas.
Para el archivo Medicion de Desempleo municipal.xlsx encontramos que tiene 143825 registros y 13 columnas.
Para el archivo Salud.xlsx encontramos que tiene 411259 registros y 13 columnas.
Para el archivo Vivienda y Servicios Publicos.xlsx encontramos que tiene 220662 registros y 13 columnas.
Para el archivo TerriData_Dim3_viviendaServiciosPublicos.xlsx encontramos que tiene 220662 registros y 13 columnas.
Para el archivo TerriData_Dim4_Educacion.xlsx encontramos que tiene 294124 registros y 13 columnas.
Para el archivo TerriData_Dim5_salud.xlsx encontramos que tiene 469609 registros y 13 columnas.
Para el archivo TerriData_Dim14_pobreza.xlsx encontramos que tiene 20040 registros y 13 columnas.

4. Implementación de los ETL's

Infraestructura:



Para comenzar este proceso fue necesario crear una conexión en Airflow. En esta conexión se hace referencia a la base de datos anteriormente creada en postgres para que se puedan acceder a estos datos.

Acto seguido se comenzó la implementación de los dags. Para esto fue necesario crear 5 archivos python en la carpeta dags dentro de la carpeta "Airflow-Docker" mencionada anteriormente. Se creo una carpeta utils en la cual se crearon 3 archivos:

1. crear_tablas.py: para el contenido de este archivo solo fue necesario copiar la información con las sentencias utilizadas para la creación de las tablas en la base de datos de postgres, no fue necesario hacerle ninguna modificación.
2. file_util.py: este archivo es el encargado del manejo de los datos, la descarga, limpieza y carga de estos. Para crearlo fue necesario copiar el código que ya nos proporcionaba el tutorial del laboratorio 5 que incluía las funciones para hacer la carga de los archivos csv.
3. insert_queries.py: este archivo es encargado de crear las sentencias utilizadas para la carga de los datos del csv a la base de datos. Reutilizamos el código del archivo del lab 5, y cambiamos los valores de cada tabla para que correspondieran a los datos de este proyecto.

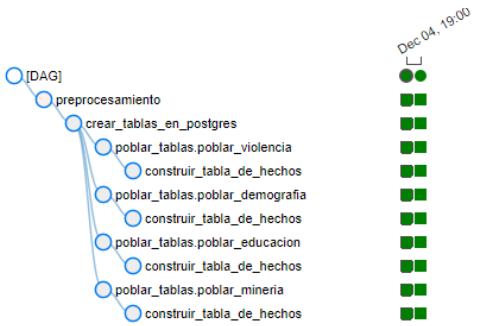
Ya para la implementación del DAG se creó dos archivos en la carpeta dags llamados ETL_mujeres.py y ETL_violencia.py, e.e archivo es el encargado de manejar el proceso del dag. Se encarga de hacer llamado a las funciones implementadas en los archivos dentro de la carpeta utils para el manejo de los datos, creación de la base de datos y poblar todas las tablas. Este archivo fue modificado del archivo del laboratorio 5 para cumplir con los datos de este nuevo proyecto.

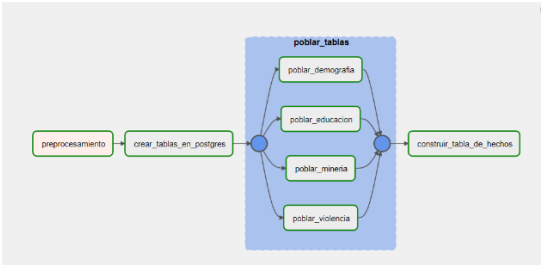
Después de tener todos los archivos necesarios para la ejecución del DAG se ejecutó este en la interfaz de Airflow. La primera vez que se ejecutó aún no se habían hecho todas las modificaciones a las sentencias para que los datos ingresados correspondieran a el tipo de dato de la columna de la tabla en la base de datos, es por esto que al principio la ejecución no fue exitosa. Nos dimos cuenta de nuestros errores a través de los logs generados por Airflow e hicimos las

modificaciones necesarias. Cuando se volvió a ejecutar el proceso se generó un nuevo error debido a que no se eliminaron los registros en las bases de datos, y los registros que se estaban cargando contenían llaves primarias ya utilizadas, y por ende, no podían ser cargados. Para solucionar este problema se eliminaron los datos a través de sentencias sql en la base de datos (este archivo se encuentra adjunto en la entrega). Ya cuando se solucionó este problema se pudo ejecutar el dag de manera exitosa.

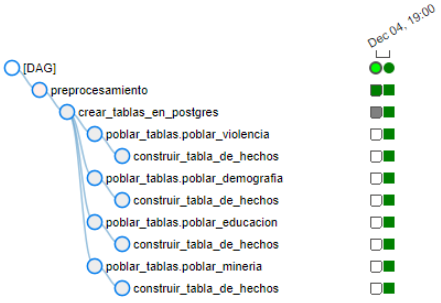
5. Resultados

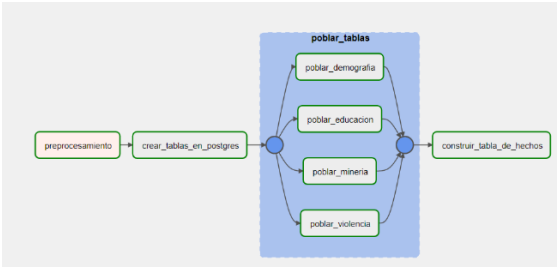
a. Mujeres:





b. Violencia:





Describir las actividades realizadas

Nuestro equipo está conformado por tres integrantes:

- María Paula González Escallón:
    - María Paula es la encargada liderar el proyecto. Se encarga de manejar las preentregas de cada miembro del equipo, asignar tareas y de resolver conflictos. También ella hace parte del liderazgo de analítica, ya que fue la encargada de verificar que la limpieza y los datos limpios fueran los adecuados para cada dimensión y así obtener los mejores resultados. También hizo parte de la realización de los ETL's, hizo parte de su diseño e implementación.
    - Las horas de trabajo fueron 12.
    - Debido a su colaboración en el equipo y a las tareas realizadas se ha decidido que a María Paula se le asignan 33,33 de los 100 puntos en total.
- Jessica A. Robles Moreno:
    - Jessica es la líder del negocio, la encargada de verificar que se cumplan con las necesidades del negocio, que los modelos dimensionales escogidos sean los que mejor respondan a las necesidades del cliente. También hizo parte de la realización de los ETL's, hizo parte de su diseño e implementación.
    - Las horas de trabajo fueron 12.
    - Debido a su colaboración en el equipo y a las tareas realizadas se ha decidido que a Jessica se le asignan 33,33 de los 100 puntos en total.
- Juan Esteban Vergara
    - Juan es el encargado de liderar el manejo de datos. El hizo la limpieza de estos para asegurar que los datos iban a ser entendidos por los modelos y así dar mejores resultados. Hizo parte del diseño y creacion de los tableros de control.
    - Las horas trabajadas fueron 12.
    - Debido a su colaboración en el equipo y a las tareas realizadas se ha decidido que a Juan Esteban se le asignan 33,33 de los 100 puntos en total.