Proyecto de Integración y Automatización de Datos para IA para el proyecto de "Diseño e implementación de un mecanismo ETL que permita la integración de fuentes públicas del gobierno local para estudios de siniestralidad vial en Bogotá"

Natalia Garzón Arias[1]

Universidad Central
Maestría en Analítica de Datos
Curso de Automatización e Integración de datos
Bogotá, Colombia
ngarzona@ucentral.edu.co

December 1, 2023

Contents

1 Introducción

Para Colombia, la reducción de muertes y lesionados graves derivados de accidentes viales ha sido una prioridad. Muestra de ello es la creación y publicación de la Ley 1702 por la cual se crea la Agencia Nacional de Seguridad Vial (Congreso de Colombia, 2013). Sin embargo, estos esfuerzos del gobierno local y nacional no parecen coincidir con las alarmantes cifras de siniestralidad en las vías.

De acuerdo con estadísticas ofrecidas por el Observatorio Nacional de Seguridad Vial, de la ANSV, en los últimos 5 años las víctimas fatales han aumentado en un 27% a nivel nacional y un 11% para Bogotá. Esto confirma que sigue siendo un reto importante la identificación de zonas con mayor riesgo de accidentalidad y la caracterización de los factores que contribuyen a su ocurrencia dentro de la ciudad (Agencia Nacional de Seguridad Vial [ANSV], s.f.).

Esta preocupación da origen al proyecto de investigación que recibe como título: "Implementation of an intelligent transportation system (ITS) for the prediction of traffic accident risk in Bogotá D.C., Colombia", en adelante nombrado proyecto ITS, con el cual sus investigadores, buscan desarrollar un prototipo de un sistema de transporte inteligente ITS capaz de predecir y mostrar en tiempo real los niveles de riesgo de accidentes de tránsito en la ciudad de Bogotá, Colombia, a través de la ubicación geográfica.

De esta manera, el presente trabajo de grado será enfocado en la participación de dos de las fases mencionadas en el proyecto ITS: (1) la recopilación de datos de las bases de datos de las instituciones del gobierno local y (2) la construcción del dataset a través de la integración y preprocesamiento de los datos recopilados. Con este planteamiento se dará respuesta a la pregunta: ¿En qué condiciones es posible diseñar e implementar un mecanismo ETL que permita la integración de fuentes públicas, heterogéneas y dispuestas por el gobierno local para considerarse como insumo de estudios de siniestralidad vial en Bogotá?

Para abordar la primera fase, se propone integrar datos de acceso público y disponibles a través de diferentes portales web, por mencionar algunos de ellos serían: la Secretaría Distrital de Movilidad (SDM) [**?**], la Infraestructura de Datos Espaciales para el Distrito Capital (IDECA) y la Dirección de Tránsito y Transporte de la Policía Nacional.

Para la segunda fase, se propone desarrollar un mecanismo bajo una política de gobierno de datos con lineamientos claros para la adquisición, tratamiento y almacenamiento de los datos, asegurando su calidad, consistencia y confidencialidad de manera que puedan estar disponibles para estudios posteriores tales como: construcción de modelos predictivos de accidentes de tráfico, identificación de puntos calientes en la ciudad y cualquier tipo de análisis que facilite la toma de decisiones en materia de seguridad vial en Bogotá. Así como la generación

de estrategias que promuevan la reducción de víctimas en las vías derivado de los datos encontrados, pues como lo demuestran las cifras, durante los meses de enero y febrero del año 2023, los accidentes de tráfico en Colombia han resultado en la muerte de 1.317 personas, lo que indica un aumento del 19,1% con respecto al promedio de los últimos cinco años (Nova, D., 2023).

Por tanto, se puede afirmar que con el desarrollo del presente trabajo se aporta al logro de los objetivos del proyecto ITS, al permitir la integración de fuentes de datos relevantes para el estudio de la siniestralidad vial en Bogotá.

En las siguientes páginas el lector encontrará la metodología utilizada para la consecución de los objetivos trazados, así como los resultados y consideraciones de este trabajo, destacando su contribución al avance del conocimiento en el campo de la analítica de datos aplicada a la seguridad vial.

2 Características del proyecto de investigación que hace uso de Integración y Automatización de Datos para IA

La industria 4.0, donde la analítica sigue jugando un papel fundamental, hoy permea todos los ámbitos, incluidos conceptos sociales y políticos como lo es la salud pública. Pues de acuerdo con la Organización Mundial de la Salud (OMS) los accidentes viales son considerados una de las causas principales de mortalidad en jóvenes. Es entonces como el presente trabajo de grado, propone una iniciativa basada en datos para abordar la problemática de la siniestralidad vial en Bogotá, Colombia y así facilitar la toma de decisiones en esta materia. Pues el propósito se centra en la creación de un mecanismo de Extracción, Transformación y Carga (ETL) para la obtención de información proveniente de diversas fuentes públicas. En la actualidad diferentes organismos gubernamentales han enfocado sus esfuerzos en recolectar, almacenar y poner a disposición de la comunidad información relacionada con la seguridad vial. Ejemplo de ello es el Observatorio Nacional de Seguridad Vial (ONSV), quien se encarga del manejo y la gestión de la información y el conocimiento relacionado con la seguridad vial en el país quien a través de diferentes elementos expone las estadísticas de siniestralidad vial. Por otra parte la Secretaría Distrital de Movilidad, a través de su Sistema Integrado de Información sobre Movilidad Urbana Regional (SIMUR) pone a disposición información asociada al comportamiento de la movilidad en Bogotá. Esto deja en evidencia que no es la falta de información lo que hace que la toma de decisiones no se haga fundamentada en datos, sino que los datos se encuentran aislados dificultando ver las relaciones que unas variables presentan sobre otras. Por esto, establecer un sistema ETL eficiente y adaptable que abarque la recopilación de datos provenientes de fuentes gubernamentales, seguida por la creación de un conjunto de datos coherente mediante procesos de integración y preprocesamiento a través de algoritmos

de limpieza y transformación de datos para asegurar su calidad y consistencia contribuye al análisis de siniestralidad en Bogotá.

2.1 Titulo del proyecto de investigación

Diseño e implementación de un mecanismo ETL que permita la integración de fuentes públicas del gobierno local para estudios de siniestralidad vial en Bogotá.

2.2 Objetivo general

Diseñar e implementar un mecanismo ETL (Extracción, Transformación y Carga) que, apoyado en una política de gobierno de datos, permita la integración de datos de fuentes públicas, heterogéneas y dispuestas por el gobierno local para que puedan ser insumo de estudios de siniestralidad vial en Bogotá.

2.2.1 Objetivos especificos

- Identificar fuentes de datos públicas que puedan contribuir al análisis de siniestralidad en Bogotá, a través de una distinción detallada de su estructura y calidad de la información contenida.
- Desarrollar un mecanismo de extracción de datos que permita obtener la información relevante de las fuentes identificadas y transformarla en un formato adecuado para su integración en el modelo de datos diseñado.
- Definir una política de gobierno de datos que permita establecer los criterios y estándares para la integración de las fuentes de datos heterogéneas en el proyecto, asegurando la consistencia y precisión de la información.
- Implementar las reglas de negocio definidas, garantizando así un proceso de transformación (limpieza, validación, creación, entre otros) de datos que asegure calidad y coherencia de la información integrada.
- Disponer los datos integrados a través de una plataforma que ofrezca información confiable y disponible para ser insumo de estudios de siniestralidad vial en Bogotá.

2.3 Alcance

Al llevar a cabo la propuesta de desarrollo de la ETL, se espera obtener una herramienta que permita integrar y transformar los datos de fuentes heterogéneas y públicas para obtener información útil y valiosa que soporte la toma de decisiones en el ámbito de la movilidad urbana y la seguridad vial. Entre los resultados concretos que se esperan obtener al implementar la ETL se incluyen:

- Aportar al logro de los objetivos del proyecto ITS: al permitir la integración de fuentes de datos relevantes para el estudio de la siniestralidad vial en Bogotá, iniciando con la recopilación de datos de las bases de datos

de las instituciones del gobierno local y concluyendo con la construcción del dataset a través de la integración y preprocesamiento de los datos recopilados en la fase previamente descrita.

- Mejora en la eficiencia y productividad: Al automatizar el proceso de integración de datos, se pueden ahorrar muchas horas de trabajo manual. Esto permite que los analistas de datos y otros profesionales puedan dedicar más tiempo a analizar los datos y tomar decisiones informadas, en lugar de destinar tiempo en tareas repetitivas y tediosas.

- Identificación de patrones y tendencias: al analizar los datos de movilidad urbana y seguridad vial desde diferentes perspectivas, se podrían identificar patrones y tendencias que permitan comprender mejor el comportamiento de los usuarios de la vía, las causas de los accidentes y otros aspectos relevantes para la toma de decisiones.

- Mejora de la toma de decisiones: al contar con información actualizada y detallada sobre el comportamiento de los usuarios del transporte público y la seguridad vial en la ciudad, se podrían tomar decisiones más informadas y eficaces en cuanto a políticas públicas de transporte y seguridad vial.

- Optimización de recursos: con la información proporcionada por la ETL, se podrían identificar áreas de la ciudad que requieren más recursos en términos de transporte público y seguridad vial, lo que permitiría una mejor adquisición de recursos y una mayor eficiencia en la planificación urbana.

- Mejora de la seguridad vial: al analizar los datos de accidentes de tráfico, se podrían identificar las áreas de la ciudad con mayor riesgo.

2.4 Pregunta de investigación

¿En qué condiciones es posible diseñar e implementar un mecanismo ETL que permita la integración de fuentes públicas, heterogéneas y dispuestas por el gobierno local para considerarse como insumo de estudios de siniestralidad vial en Bogotá?

2.5 Hipotesis

A pesar de que el gobierno local recopila, almacena y pone a disposición datos relacionados con accidentes de tráfico y siniestralidad vial a través de diferentes fuentes de acceso público, como el Instituto de Desarrollo Urbano de Bogotá (IDU), la Secretaría Distrital de Movilidad (SDM), la Infraestructura de Datos Espaciales para el Distrito Capital (IDECA) y la Dirección de Tránsito y Transporte de la Policía Nacional, existe un desafío significativo en la integración y consolidación de estos datos dispersos en un formato coherente y accesible. La falta de un mecanismo robusto y eficiente de Extracción, Transformación y Carga (ETL)

dificulta la obtención de información completa y precisa para llevar a cabo estudios detallados sobre la siniestralidad vial en Bogotá.

El proceso de recopilación, limpieza, transformación e integración de datos de diferentes fuentes es esencial para realizar análisis precisos y obtener información valiosa para la toma de decisiones en políticas de seguridad vial. Sin embargo, la ausencia de un mecanismo ETL especializado que permita la integración de las diversas fuentes públicas de datos del gobierno local limita la capacidad de investigadores, planificadores urbanos y responsables de políticas para comprender los patrones de siniestralidad vial, identificar áreas críticas y diseñar estrategias efectivas de prevención.

Por lo tanto, surge la necesidad de abordar este problema mediante la implementación de un mecanismo ETL que facilite la recopilación, transformación e integración de los datos de siniestralidad vial dispersos en diferentes fuentes públicas proporcionadas por el gobierno local de Bogotá. Este mecanismo debe ser capaz de estructurar los datos de manera coherente, eliminar duplicados, tratar inconsistencias y generar una base de datos unificada que permita un análisis más profundo y preciso de la siniestralidad vial en la ciudad.

En este contexto, la presente investigación tiene como objetivo diseñar e implementar un mecanismo ETL (Extracción, Transformación y Carga) que, apoyado en una política de gobierno de datos, aborde la problemática de la integración de datos de fuentes públicas, heterogéneas y dispuestas por el gobierno local para que puedan ser insumo de estudios de siniestralidad vial en Bogotá. El propósito de esta investigación es mejorar la disponibilidad y calidad de los datos utilizados en la toma de decisiones en materia de seguridad vial y planificación urbana.

3 Reflexiones sobre el origen de datos e información

3.1 ¿Cuál es el origen de los datos e información?

Los datos se encuentran distribuidos en diferentes repositorios y publicados a través del portal de datos abiertos de la Secretaría Distrital de Movilidad (SDM).

3.2 ¿Cuáles son las consideraciones legales o éticas del uso de la información?

Los datos abiertos son información pública dispuesta en formatos que permiten su uso y reutilización bajo licencia abierta y sin restricciones legales para su aprovechamiento. En Colombia, la Ley 1712 de 2014 sobre Transparencia y Acceso a la Información Pública Nacional, define los datos abiertos en el numeral sexto como "todos aquellos datos primarios o sin procesar, que se encuentran en formatos estándar e interoperables que facilitan su acceso y reutilización, los cuales están bajo la custodia de las entidades públicas o privadas que cumplen con funciones públicas y que son puestos a disposición de cualquier ciudadano, de forma libre y sin restricciones, con el fin de que terceros puedan reutilizarlos y crear servicios derivados de los mismos". De este modo, la Ley establece la obligatoriedad de las entidades públicas de "divulgar datos abiertos", teniendo en cuenta las excepciones de acceso a la información, asociadas a información clasificada y reservada establecidas en su título tercero, artículos 18 y 19.

3.3 ¿Cuáles son los retos de la información y los datos que utilizara en Integración y Automatización de Datos para IA?

El poyecto enfrenta los siguientes desafíos:

- **1. Calidad de datos:** Los datos pueden estar incompletos, duplicados o incorrectos. Limpiar y transformar estos datos de manera efectiva es el desafío crítico en este proceso ETL.
- **2. Variedad de datos:** La diversidad de fuentes con recopilación de información similar podría generar redundancia y discrepancia de los datos desde diferentes fuentes
- **3. Gestión de errores:** Los errores pueden ocurrir en cualquier etapa del proceso ETL. Diseñar mecanismos de detección y manejo de errores eficaces es esencial.
- **4. Monitorización y mantenimiento:** La monitorización constante de los flujos de datos y el rendimiento del proceso ETL es crucial para identificar problemas y mantener la integridad de los datos.
- **5. Documentación y trazabilidad:** Mantener un registro detallado de las transformaciones y cambios realizados en los datos es importante para la trazabilidad y la auditoría.

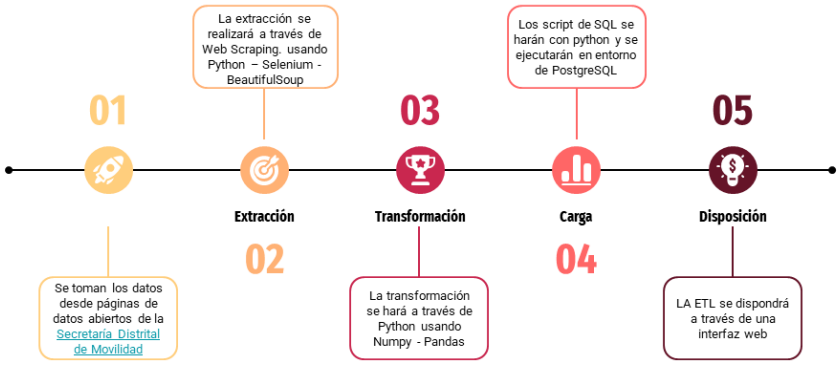
3.4 ¿Que espera de la utilización de Integración y Automatización de Datos para IA para su proyecto?

Se espera que se libere la carga operativa de descargar, transformar y poner a disposición para su posterior utilización la información desde diferentes platafomas públicas. La obtención de los datos se realizará por medio de un proceso de *web scraping*, a través de Python y haciendo uso de la librería *Selenium*. Esta librería permitirá la automatización web emulando la interacción humana en un navegador. Las transformaciones para convertir, limpiar y enriquecer los datos provenientes de las diversas fuentes extraídas en la fase anterior se realizarán igualmente a través Python y haciendo uso de las librerías *Numpy y Pandas*. Finalmente para la carga se hace la creación de los archivos de sql a través de Python y se llevan a PostgreSQL para ser ejecutados desde Pg Admin o Shell.

4 Diseño de integración y Automatización de Datos para IA (Diagrama)

En la figura 1 se detalla el diagrama que orientará el proyecto en la automatización de extracción, transformación y carga de datos para crear una base de datos de la que se pueda extraer información relevante para analizar la siniestralidad en la ciudad de Bogotá.

Figure 1: Diagrama proyecto ETL



Nota. Conozca los datos abiertos de la Secretaría Distrital de Movilidad en el enlace disponible en el pie de imagen. [a]

[a] `https://datos.movilidadbogota.gov.co/search?groupIds=d3812f8315054cdc84cf744680103713`

5 Integración de Datos

Implementación del Proceso ETL con Python

El proceso ETL se desarrolló mediante el uso de lenguaje de programación Python. Estos scripts se diseñaron para extraer los archivos en formato .csv desde los repositorios web, llevar a cabo las transformaciones necesarias en los datos adquiridos y, posteriormente, cargarlos en la base de datos PostgreSQL.

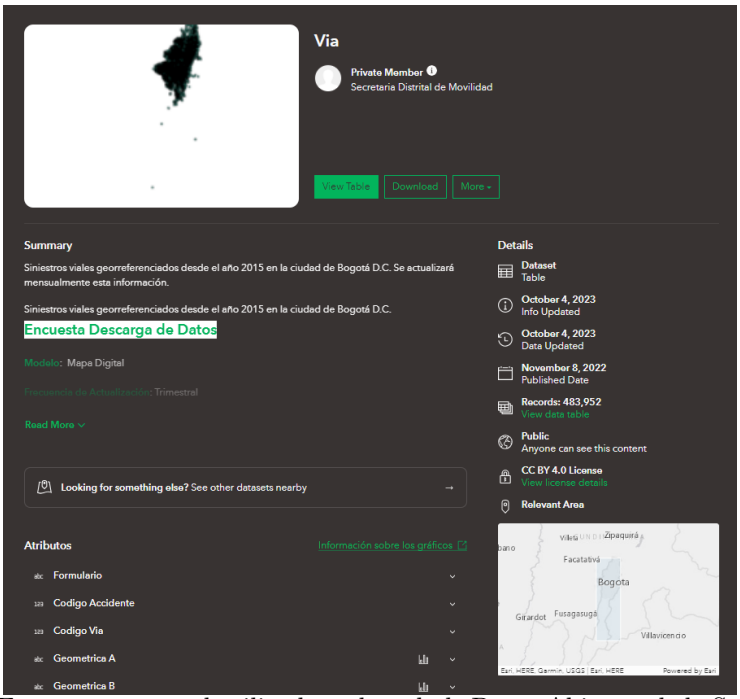
5.1 Extracción de Datos

Como se ha mencionado a lo largo del documento, la obtención de los datos se realizó por medio de un proceso de *web scraping*, a través de Python y haciendo uso de la librería *Selenium*. Esta librería permitió la automatización

web emulando la interacción humana en un navegador. De manera muy general, el proceso consistió en el reconocimiento y navegación de los sitios web para conocer y entender su estructura y la descarga de los conjuntos de datos relevantes. A continuación, se describen con mayor detalle los pasos realizados en esta etapa de extracción. Los algoritmos creados en Python podrán ser consultados en detalle en el anexo C:

1. Reconocimiento y navegación de los sitios web: Las páginas web que contienen los conjuntos de datos desacargados[1] manejan una estructura similar en su diseño. Como se puede apreciar en la Figura 2, correspondiente al conjunto de datos de vía, la página se compone de dos elementos esenciales: (1) textos con información asociada al *dataset* como: resumen, atributos, fechas de publicación y actualización, entre otros. Y (2) botones de interacción para descargar o ver los datos en formato tabla.

Figure 2: Estructura general de las páginas desde las cuales se extrajeron los datos



Nota. Estructura general utilizada en la web de Datos Abiertos de la Secretaría Distrital de Movilidad para sus diferentes conjuntos de datos.

El proceso de exploración web inicia de forma manual desde el navegador,

[1] `https://datos.movilidadbogota.gov.co/search?groupIds=d3812f8315054cdc84cf744680103713`

seguido por el reconocimiento del código HTML desde Python a través de las librerías tradicionales para realizar *scraping*, como *MechanicalSoup* y *BeautifulSoup*. Esta exploración inicial a través de un lenguaje de programación, permitió identificar que una gran parte del código HTML era generado a través de scripts de JavaScript que interactuaban con elementos en el *Document Object Model* (DOM) para cambiar su contenido. Razón por la cual se descartan las herramientas anteriormente descritas y se emplean herramientas que permitan interactuar con el sitio web

2. Utilización de un navegador web automatizado: Atendiendo a la limitación descrita en el numeral 1, se utilizó un navegador web automatizado mediante Python para acceder a las páginas web que alojaban los conjuntos de datos. Para lograr esto se hizo uso de la librería *Selenium* y de su controlador *(WebDriver)*, asegurando algunos pasos previos:

    - se seleccionó Chrome como navegador.
    - se instalaron las dependencias para instalar el navegador seleccionado
    - se instaló Google-Chrome
    - se descargó el *WebDriver* para Chrome
    - se importó el módulo *WebDriver* de *Selenium*
    - se creó un objeto *Service* para especificar la ruta del controlador *(WebDriver)*
    - se creó una instancia del navegador Chrome

3. Interacción con las páginas web: Se exploararon dos alternativas para la obtención de los datos:

    - mostrar los datos en formato tabla y crear un *dataset* a partir del recorrido de una tabla de *scroll* infinito
    - hacer clic en el botón descargar para obtener el archivo en el formato deseado

    Se seleccionó la opción de descarga, para lo cual se programó la interacción con el botón *Download* para desencadenar la descarga del archivo. En esta interacción se identifican la presencia de árboles DOM ocultos a los elementos en el árbol DOM principal.

4. Navegación en elementos sombra: Se definió una función que permitiera acceder a los elementos ocultos y así poder navegar, en cascada, a través de dichos elementos para acceder al botón de descarga y enseguida al botón del formato seleccionado, que para el mecanismo ETL se eligieron los archivos en formato .CSV.

5.2 Transformación de los Datos

En las siguientes líneas se describen el conjunto de transformaciones diseñadas para convertir, limpiar y enriquecer los datos provenientes de las diversas fuentes extraídas en la fase anterior.

- **Revisión de la estructura:** Aunque este ítem no cumple propiamente alguna de las acciones delimitadas en la transformación, prepara el terreno para hacerlo. Pues la intención de este primer momento es conocer cómo se encuentra distribuida la información e identificar los formatos en los que se almacenan los datos en las diferentes variables.
- **Exploración de los datos:** Siguiendo con la preparación de los datos se calculan las estadísticas descriptivas tanto para variables numéricas como categóricas.
- **Contabilización de valores nulos:** Se identifican los valores nulos así como su participación en el total de registros por columnas. La intención de este paso es eliminar aquellos que tengan una alta participación y que por el aporte que darían al set de datos se esperaría que fuera muy baja.
- **Identificar registros duplicados:** Este paso contemplado dentro del proceso e limpieza, busca eliminar registros que puedan aumentar las cifras, entregando una interpretación errónea.
- **Identificar uniformidad de elementos dentro de las variables:** Con este ítem se busca identificar variedad en la escritura de los elementos en las variables en las que se espera exista uniformidad.
- **Filtrado de datos:** En este apartado se seleccionan únicamente las variables que se esperan llevar al paso de carga.
- **Asegurar el formato de los datos:** En este paso se convierten las variables que lo requieran al tipo de datos que se esperan poblar en la base de datos empleada en la etapa de carga.
- **Creación de características nuevas:** Con el ánimo de alivianar la tabla de hechos se crean las dimensiones de tiempo, gravedad, clase de accidente y localidad.
- **Codificación de variables categóricas**: Con la creación de las dimensiones descritas en el ítem anterior, en la tabla de hechos se transforman los atributos tiempo, gravedad, clase de accidente y localidad a sus equivalentes en códigos.
- **Eliminación de datos incongruentes:** Esta categoría hace referencia de manera particular al atributo fecha de nacimiento de la tabla Actor Vial. Se reemplazan por valores nulos todas las fechas que por interpretación del software se leen como 01-ene-1900 y todas aquellas que sus años estén por encima del año de la fecha en la que se ejecuta el *notebook*.

Estas transformaciones fueron cruciales para asegurar que los datos fueran coherentes y relevantes para la sieguiente etpa. La etapa de carga.

5.3 Carga:

Previo a la fase de carga se debe disponer de un escenario en el cual almacenar los datos extraídos, transformados y limpiados. Bajo este contexto la fase de carga contemplará: (1) La creación de la base de datos con las tablas correspondientes y (2) el proceso de poblamiento de las tablas previamente creadas.

5.3.1 Creación de la Base de Datos y sus Tablas

El proceso de creación de la base de datos inició con el **modelado de datos**, siguiendo un enfoque relacional para organizar la información de manera eficiente. A continuación, se describen las tablas principales y sus atributos:

- Descripción del Esquema de Base de Datos:

  Se detalla a continuación el esquema de base de datos generado por la herramienta ERD en pgAdmin 4. Este esquema constituye una parte esencial de un sistema diseñado para la gestión y análisis de incidentes de tráfico.

- Tablas Representando Entidades Relevantes:

  - `agente_transito`
  - `calzadas`
  - `carriles`
  - `clase_accidente`
  - `clase_vh`
  - `condicion_actor`
  - `condiciones`
  - `describe_causa`
  - `estado`
  - `genero`
  - `geometrica_a`
  - `geometrica_b`
  - `geometrica_c`
  - `gravedad`
  - `horario`
  - `iluminacion_a`
  - `localidad`

- material
- modalidad
- semaforo
- servicio
- tiempo
- tipo_causa
- utilizacion

Tablas Relacionadas con Incidentes y Actores Involucrados:

- accidente
- actor_vial
- causa
- vehiculo
- via

Relaciones Definidas por Claves Externas:

- La tabla `accidente` establece claves externas que hacen referencia a otras tablas para diversos atributos como `cod_clase_acc`, `cod_gravedad`, `cod_localidad`, entre otros.
- La tabla `actor_vial` define claves externas que hacen referencia a las tablas `estado`, `genero`, `condicion_actor` y `accidente`.
- La tabla `causa` presenta claves externas que hacen referencia a `tipo_causa`, `describe_causa` y `accidente`.
- La tabla `vehiculo` incorpora claves externas que hacen referencia a `clase_vh`, `modalidad`, `servicio` y `accidente`.
- La tabla `via` establece claves externas que hacen referencia a varias tablas como `geometrica_a`, `calzadas`, `carriles`, etc., y también hace referencia a la tabla `accidente`.

Este esquema de base de datos configura un modelo relacional con integridad referencial, proporcionando una estructura robusta para el almacenamiento y análisis de datos relacionados con incidentes de tráfico. Cualquier pregunta adicional o consulta específica relacionada con este esquema puede ser abordada de manera más detallada.

En la Figura **??**, se presentan las relaciones que poseen una importancia fundamental para comprender la manera en que los datos interactúan en la base de datos y su relevancia en el análisis de información relativa a los accidentes viales.

Como se evidencia en la figura, la mayoría de estas relaciones son de tipo "uno a muchos", lo que implica que un registro en una tabla puede estar relacionado con varios registros en otra tabla.

5.3.2 Normalización y restricciones

Para garantizar la estructura definitiva de la base de datos y mejorar la integridad de los datos transformados, se aplicaron reglas de normalización y se establecieron restricciones de integridad referencial. A continuación, se detallan las acciones tomadas en esta etapa:

Normalización La normalización de la base de datos se llevó a cabo con el propósito de eliminar redundancias y minimizar anomalías. Las acciones de normalización clave incluyen:

- **Identificación de campos clave:** Se han identificado campos clave en las tablas para garantizar la unicidad de los registros. Por ejemplo, el campo `CODIGO_ACCIDENTE` se ha definido como clave primaria en varias tablas.
- **Eliminación de redundancias:** La información se ha fragmentado en tablas relacionadas, de modo que cada tabla contenga datos específicos relacionados con un único aspecto (por ejemplo, actor vial, causa, vehículo, etc.).
- **Minimización de anomalías:** La normalización ha ayudado a reducir las anomalías relacionadas con operaciones de actualización, inserción y eliminación, al eliminar la necesidad de modificar la misma información en varios lugares.
- **Cumplimiento de formas normales:** El modelo de datos cumple con la Primera Forma Normal (1NF), asegurando que las tablas estén estructuradas de tal manera que se eliminen los valores repetidos y las listas dentro de celdas.

Restricciones de integridad referencial Se han establecido restricciones de integridad referencial para mantener la coherencia y la integridad de los datos. Las acciones clave relacionadas con las restricciones son las siguientes:

- **Establecimiento de claves primarias y foráneas:** Se han definido claves primarias y foráneas para relacionar las tablas. Por ejemplo, la tabla "Actor Vial" se relaciona con la tabla "Accidente" a través del campo `CODIGO_ACCIDENTE`.
- **Mantenimiento de integridad referencial:** Las claves foráneas garantizan que no se puedan crear relaciones entre registros inexistentes, lo que mantiene la integridad referencial.
- **Restricciones específicas:** Si se han aplicado restricciones *UNIQUE* o *CHECK* en campos particulares, se han documentado junto con su propósito.

- **Relaciones uno a muchos y uno a uno:** Se han definido y documentado las relaciones uno a muchos y uno a uno entre las tablas, como la relación de “uno a muchos” entre “Accidente” y “Actor Vial”, ya que un accidente puede tener varios actores viales asociados.

A continuación, se presentan algunos ejemplos que ilustran la aplicación de claves primarias, claves foráneas y restricciones en el modelo de datos. Estos ejemplos tienen como propósito facilitar la comprensión de la estructura de la base de datos y la implementación de las restricciones.

6 Automatización de Datos

La automatización del proceso se estimó como una automatización asistida en donde el usuario debe ejecutar los prcoesos. Se estima que la intervención del ususario se de a través de una intrfaz web, interfaz que se encuentra en proceso de desarrollo.

7 Proximos pasos

Para futuras intervenciones se espera incluir nuevas fuentes de información que enriquezcan los datos y se puedan incluir variables que hasta el momento han sido desestimadas en el análisis e interpretación de la sinietralidad vial.

8 Lecciones aprendidas

- Se adquirieron conocimientos de web scraping, entendiendo la forma de estructurar, leer e interpretar una página web
- Todos los procesos son suceptibles a ser automatizados, siempre y cuando se tenga claro un digrama de flujo del proceso. La diagramación permitió reconocer herramientas a ser incluidas dentro del proceso, que fases debian ocurrir antes y concebir en un solo panorama qué tanto ´puede ser automatizado con o sin intervención del usuario.
- Concebir el alcance de acuerdo con el horizonte de tiempo trazado

9 Bibliografía

- Abulatif, L. (2018). Data integration process -an information management framework for multiple road crash databases in brazil. Epidemiologia e Serviços de Saúde, 27 (2). Descargado de https://doi.org/10.5123/s1679-49742018000200018.
- Agencia Nacional de Seguridad Vial (ANSV). (2018). Observatorio - estadísticas. histórico de víctimas (Inf. Téc.). Agencia Nacional de Seguridad Vial (ANSV). Descargado de https://ansv.gov.co/es/observatorio/estad
- Agencia Nacional de Seguridad Vial (ANSV). (2019). Hipótesis de causas probables en los siniestors viales de colombia. (Inf. Téc.). Agencia Nacional de Seguridad Vial (ANSV).
- Aldana, R. (2021, 27 de mayo). Factores de riesgo en los accidentes de tráfico. Descargado de https://www.aulacarreteras.com/factores-riesgo-accidentes-trafico/ (Recuperado el 3 de septiembre de 2023)
- Barón Benavides, G. (2020). Análisis del agendamiento de la accidentalidad vial en el marco de la política pública de seguridad vial en bogotá (2016-2017). Descargado de https://bdigital.uexternado.edu.co/handle/001/3492
- Congreso de Colombia. (2002). Ley 769. (Por la cual se expide el Código Nacional de Tránsito Terrestre y se dictan otras disposiciones. 6 de agosto de 2002. D.O. No. 45026)
- Congreso de Colombia. (2013). Ley 1702. (Por la cual se crea la Agencia Nacional de Seguridad Vial y se dictan otras disposiciones. 27 de diciembre de 2013. D.O. No. 49016)
- Connolly, T. M., y Begg, C. E. (2014). Database systems: A practical approach to design, implementation, and management. Pearson.
- DAMA International. (2017). Dama-dmbok: Data management body of knowledge (2nd ed.).Technics Publications.
- Date, C. J. (2003). An introduction to database systems. Addison-Wesley.
- Departamento de Ingeniería Eléctrica y Electrónica de la Universidad Nacional de Colombia.(2007). Guía didactica para el buen uso de la energía (Inf. Téc.). Unidad de PlaneaciónMinero Energética UPME, Ministerio de Minas y Energía.

  Descargado de http://www.upme.gov.co/Docs/AlumbradoPublico.pdf
- Elmasri, R., y Navathe, S. B. (2019). Fundamentals of database systems. Pearson.
- Inmon, W. H. (2005). Building the data warehouse. John Wiley Sons.
- Inmon, W. H. (2016). Mastering the data warehouse lifecycle: A step-by-step guide. Wiley Publishing, Inc.

- Instituto de Recursos Mundiales (WRI). (2018). Sostenibilidad y seguridad: Visión y marco para lograr cero muertes en las vías. (Inf. Téc.). Instituto de Recursos Mundiales (WRI). Descargado de https://wrimexico.org/sites/default/files/Sostenibilidad

- Instituto Nacional de Vías (INVIAS). (2016). Manual de mantenimiento de carreteras (Inf. Téc.). Instituto Nacional de Vías, Ministerio de Transporte. Descargado de https://www.invias.gov.co/index.php/archivo-y-documentos/proyectos-de-norma/11316-manual-de-mantenimiento-de-carreteras-2016-volumen-2-especificaciones-generales/file

- Instituto Nacional de Vías (INVIAS). (2018). Glosario de manual de diseño geométrico de carreteras (Inf. Téc.). Instituto Nacional de Vías, Ministerio de Transporte. Descargado de https://www.invias.gov.co/index.php/servicios-al-ciudadano/glosario/130-glosario-manual-diseno-geometrico-carreteras

- Instituto Nacional de Vías (INVIAS). (2022). Manual de capacidad y niveles de servicio para carreteras de dos carriles (Inf. Téc.). Instituto Nacional de Vías, Ministerio de Transporte. Descargado de https://www.invias.gov.co/index.php/archivo-y-documentos/documentos-tecnicos/12869-manual-de-capacidad-y-niveles-de-servicio-para-carreteras-de-dos-carriles-tercera-version-2022/file

- Mesquitela, J., Elvas, L. B., Ferreira, J. C., y Nunes, L. (2022). Data analytics process over road accidents datamdash;a case study of lisbon city. ISPRS International Journal of Geo-Information, 11 (2). Descargado de http://dx.doi.org/10.3390/ijgi11020143.

- Nova, D. (2023). Boletín estadístico de fallecidos y lesionados por siniestros viales: Serie nacional, enero-mayo 2023 (Inf. Téc.). Agencia Nacional de Seguridad Vial (ANSV). Descargado de https://ansv.gov.co/sites/default/files/2023-07/BoletinBogot

- Organización Mundial de la Salud. (2022). Traumatismos causados por el tránsito (Inf.Téc.). Organización Mundial de la Salud (OMS). Descargado de https://www.who.int/es/news-room/fact-sheets/detail/road-traffic-injuries.

- Palacios Pachón, C. (2015). Análisis de los accidentes de tránsito en bogotá, como problemática de salud pública y su impacto en el política pública 2010 – 2013. Descargado de https://repositorio.unbosque.edu.co/handle/20.500.12495/5538.

- PostgreSQL. (2023). Postgresql - the world's most advanced open-source database (Accessed n.o 2023-09-24). Retrieved from https://www.postgresql.org/.

- Python Software Foundation. (2023). Python software foundation (Accessed n.o 2023-09-24). Retrieved from https://www.python.org/.

- Ralph Kimball, J. C. (2004). The data warehouse. etl toolkit. Wiley Publishing, Inc.