Analytica Data Solutions

TRANSFORMANDO DATOS EN DECISIONES

Proyecto de Movilidad Sostenible NYC

INTEGRANTES:

María Marcela Balzarelli
Pablo Nahuel Barchiesi Ponce
Michael Williams Martinez Chinchilla
Jorgelina Paola Lujan Ramos

Tabla De Contenido

1) Introducción

2) Objetivos

3) Cambios respecto al sprint 1
   3.1 arquitectura
   3.2 kpi

4) ETL

5) Diccionario de datos

6) Análisis Exploratorio

7) Creación del Data Warehouse (Diagrama Entidad Relación)

8) Tareas por realizar

   8.1 automatizar el dw
   8.2 carga incremental

1. INTRODUCCIÓN

Podemos observar el trabajo realizado en el segundo sprint, el cual consiste en la creación y optimización del pipeline de datos (ETL), Asimismo, se pondrá en marcha la implementación de un datawarehouse automatizado, acompañado de la carga incremental.

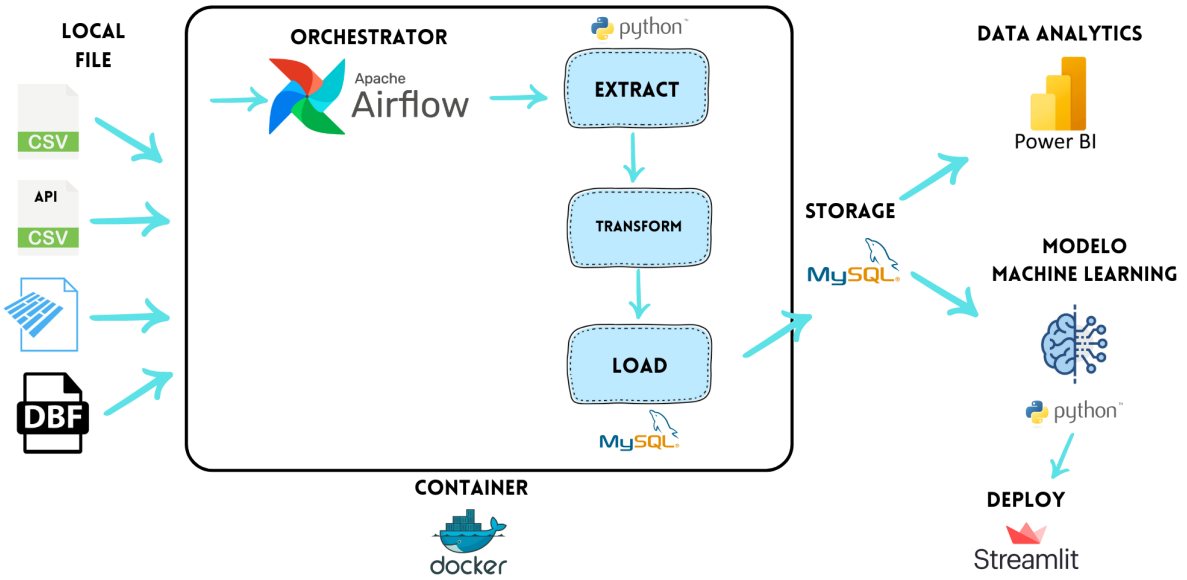
El equipo de trabajo ha demostrado una sólida capacidad de organización y colaboración para abordar las demandas del cliente. Sin embargo, se ha registrado la necesidad de introducir ajustes en la arquitectura general del proyecto, así como en la definición de los indicadores clave de rendimiento (KPIs).

2. OBJETIVOS

El objetivo principal del segundo sprint radica en la automatización completa del pipeline de datos. Esta automatización conlleva la transformación de los datos en formatos adecuados, seguida de su almacenamiento en un datawarehouse diseñado con precisión para facilitar análisis posteriores.

3. CAMBIOS REALIZADOS EN COMPARACIÓN AL SPRINT 1

3.1) Arquitectura



Ingesta de Datos:
Los archivos de datos en los formatos CSV, Parquet y DBF están almacenados localmente . Estos archivos contienen información esencial sobre la movilidad urbana y son la base de nuestro análisis.

Orquestación con Apache Airflow en Docker:
Utilizamos Apache Airflow para orquestar y programar el flujo de trabajo de ingesta, procesamiento y carga de los datos. Airflow asegura que las tareas se ejecuten automáticamente y en el orden correcto, optimizando la eficiencia del sistema.

Proceso ETL Automatizado:
El proceso de Extracción, Transformación y Carga (ETL) se realiza a través del lenguaje de programación Python, además, hemos implementado con airflow el mecanismo de carga incremental para actualizar los datos de manera eficiente.

Almacenamiento en MySQL:

Los datos procesados se almacenan en una base de datos MySQL.

Análisis de Datos en Power BI:
Una vez que los datos se encuentran en la base de datos, se pueden analizar y visualizar utilizando Power BI. Esto permite identificar tendencias, patrones y obtener información valiosa para la toma de decisiones informadas.

Modelo de Machine Learning y Streamlit:
Se creará un modelo de Machine Learning en Python. Este modelo se implementará en una aplicación interactiva utilizando Streamlit, lo que permitirá a los usuarios utilizar el modelo y obtener predicciones en tiempo real.

Justificación y beneficios de los cambios.

En Analytica Data Solutions, decidimos mantener nuestro proceso de almacenamiento y análisis de datos local en lugar de optar por una solución en la nube. Esta elección se basa en una evaluación completa de nuestras necesidades y metas. Creemos que esto nos ofrece ventajas clave:

Al mantener el proceso de datos local, tenemos control total sobre la infraestructura y recursos. Esto nos permite ajustar la configuración según nuestras necesidades, adaptándonos eficientemente a cambios en el proyecto y demandas de datos. Evitamos problemas de conexión pasados y garantizamos un flujo fluido y confiable, mejorando la productividad.

Al mantener operaciones locales, evitamos depender de un proveedor de nube específico. Esto reduce riesgos ligados a cambios en términos y precios de servicios en la nube. La independencia nos da mayor control sobre recursos y operaciones.

Mantener operaciones locales permite aprovechar la experiencia interna, aumentando la eficiencia en proyectos y la adaptación ágil a cambios.

3.2 KPI

Los siguientes KPIs se utilizarán para evaluar el éxito de nuestras soluciones y el impacto de la implementación de vehículos eléctricos en la flota de transporte:

KPI: Crecimiento Porcentual de Tarifas de Taxi Verdes dos años
Objetivo: Evaluar si hubo un aumento de al menos el 5% en las tarifas entre los dos últimos años.

KPI: Control de Incremento de Contaminación Sonora en Barrios con Mayor Contaminación
Objetivo: Limitar el incremento anual de los niveles de contaminación sonora en los barrios identificados con mayor contaminación a un máximo del 5% durante los próximos dos años.

KPI: Crecimiento Anual de la Cantidad de Viajes Realizados
Objetivo: Aumentar la cantidad de viajes realizados en un 10% anual durante los próximos dos años.

KPI: Días de la Semana con Mayor Demanda en Épocas de Bajas Temperaturas
Objetivo: Identificar los días de la semana con la mayor cantidad de viajes realizados durante los períodos de bajas temperaturas, buscando incrementar la cantidad de viajes en esos días en al menos un 10% en el próximo año.

4) ETL

De los 8 dataset entregados por la empresa se escogieron tres que son Electric and Alternative Fuel Charging Stations.csv ,en el cual se realizaron transformaciones como limpieza de nulos y se redujo a utilizar el 30% de los datos ya que son los prodigios para nuestros análisis y modelo de machine learning y se le cambió el nombre a Station_NY.csv.

taxi_zones.dbf y taxi+_zone_lookup.csv se realizó un merge entre los dos dataset para formar un dataset llamado taxis zonas.csv y el cual se eliminaron algunas columnas irrelevantes para nuestro estudio.

Los otros 5 dataset se excluyeron porque tenían muchos valores nulos como es el caso del “ alternative fuel vehicles us.csv” “ light duty vehicles.csv ” y los demás no van con nuestro caso de estudio por lo que encontramos y se extrajeron dataset de fuentes externas.

Los dataset extraidos fueron calidad_del aire.csv , conta_sonora.csv, NYCclima.csv y vehiculos_combustion_co2_2023 y un conjunto de datos taxis.parquet_2022_2023 que representan las rutas de los taxis amarillos y verdes que transitan en nueva york en el cual este ultimo se realizaron transformaciones como cambiar tipo de datos, filtro por dia, creacion de columnas nuevas, cambio de nombre a español, eliminacion de valores nulos por ultimo se escogieron las columnas a utilizar y se concatenaron los dataframe de taxis_amarillos y taxis_verdes para posteriormente concatenar estos dos ultimos en el cual se llego al taxis_tarifa.csv

Los dataset dados por la empresa y extraídos se encuentran el la carpeta llamada Datasets de nuestro repositorio y los datasets ya limpios en el cual se utilizaran para nuestro análisis y modelo de m.l se encuentran dentro de sprint_2 llamada la carpeta dataset_limpios.

5. Diccionario de datos

5.1. Conta_Sonora

Este conjunto de datos contiene registros de mediciones de sonidos recopilados en diferentes fechas y ubicaciones.

Cada registro incluye información sobre la fecha de la medición, el identificador del distrito (borough), la cantidad

de sonidos de motor detectados, la cantidad de sonidos de señales de alerta (bocina), el total de sonidos recopilados y el

nombre del distrito correspondiente.

- fecha : Indica la fecha de la medicion
- id_borough : Indica el id del barrio de la medición

- engine_sounds : Sonido del motor
- alert_signal_sounds : Señales de alerta (bocina)
- total_sounds : Total de sonidos
- borough_name : Nombre del barrio

5.2. Station_NY

Dataset referido a las estaciones de servicio, los tipos de combustibles y sus ubicaciones.

- ID: Identificador único o número de identificación para los registros. Puede ser un número secuencial o una clave única que distingue cada registro en el conjunto de datos.
- Fuel Type Code: Código que representa el tipo de combustible asociado con la estación. Puede indicar si la estación es para vehículos eléctricos, vehículos a gas natural, etc.
- Station Name: Nombre de la estación donde se encuentra el punto de recarga o abastecimiento de combustible.
- City: Ciudad donde está ubicada la estación.
- State: Estado donde está ubicada la estación.
- EV Level1 EVSE Num: Número de puntos de recarga de vehículos eléctricos (EVSE) de nivel 1 en la estación.
- EV Level2 EVSE Num: Número de puntos de recarga de vehículos eléctricos (EVSE) de nivel 2 en la estación.
- EV DC Fast Count: Cantidad de cargadores de corriente continua (DC Fast) para vehículos eléctricos en la estación.
- EV Network: Red de recarga de vehículos eléctricos asociada con la estación.
- Geocode Status: Estado del proceso de geocodificación para determinar la ubicación de la estación.
- Latitude: Latitud geográfica de la ubicación de la estación.
- Longitude: Longitud geográfica de la ubicación de la estación.
- NG Vehicle Class: Clase de vehículos a gas natural.
- EV Connector Types: Tipos de conectores utilizados para la recarga de vehículos eléctricos en la estación.
- Groups With Access Code (French): Grupos con código de acceso (en francés) que pueden utilizar la estación.
- Access Detail Code: Código que detalla el acceso a la estación.

- CNG Dispenser Num: Número de dispensadores de gas natural comprimido (CNG) en la estación.
- CNG Vehicle Class: Clase de vehículos a gas natural comprimido (CNG).
- LNG Vehicle Class: Clase de vehículos a gas natural licuado (LNG).

5.3. taxis_zone

Este conjunto de datos contiene información geoespacial y de ubicación relacionada con las zonas y áreas de la ciudad
de Nueva York. Cada registro en el conjunto de datos incluye datos como la longitud y el área de la forma geográfica,
un identificador de ubicación, el distrito (borough), el nombre de la zona y la zona de servicio.

- Shape_Leng: Longitud de la forma geográfica o contorno de una zona o área.
- Shape_Area: Área de la forma geográfica de una zona o área.
- LocationID: Identificador numérico único asignado a cada ubicación o zona.
- Borough: Nombre del distrito o barrio al que pertenece la zona.
- Zone: Nombre de la zona geográfica.
- service_zone: Nombre de la zona de servicio asociada con la ubicación.
- Latitude: Latitud
- Longitude: Longitud

5.4. Vehículos_combutión_CO2

Este conjunto de datos contiene información detallada sobre diferentes modelos de vehículos. Cada registro
incluye atributos clave como el año de fabricación, el fabricante (Make), el modelo específico (Model.1), la clase de vehículo,
el tamaño del motor en litros, el número de cilindros, el tipo de transmisión, el tipo de combustible, el consumo de combustible
en ciudad, las emisiones de CO2, la clasificación de emisiones de CO2 y la clasificación de emisiones de smog.

- Model(Year): Año de fabricación del modelo del vehículo.
- Make: Fabricante del vehículo.
- Model.1: Modelo específico del vehículo.
- Vehicle Class: Clase o categoría del vehículo.
- Engine Size(L): Tamaño del motor del vehículo en litros.
- Cylinders: Número de cilindros en el motor del vehículo.
- Transmission: Tipo de transmisión del vehículo.
- Fuel(Type): Tipo de combustible utilizado por el vehículo.
- Fuel Consumption(City (L/100 km): Consumo de combustible en ciudad en litros por cada 100 kilómetros recorridos.
- CO2 Emissions(g/km): Emisiones de dióxido de carbono del vehículo en gramos por kilómetro.
- CO2(Rating): Clasificación de emisiones de dióxido de carbono del vehículo.
- Smog(Rating): Clasificación de emisiones de smog del vehículo.

5.5. TaxiG

Este conjunto de datos contiene información sobre transacciones de recogida de taxis eléctricos en un período específico. Cada registro incluye detalles como la fecha de recogida , el día de la semana , el monto total de la transacciones y la cantidad de transacciones registradas.

- pickup_date: Fecha en la que se realizó la recogida de taxi.
- weekday: Día de la semana en el que tuvo lugar la recogida de taxi.
- total_amount: Monto total de las transacciones correspondiente a la fecha.
- row_count: Cantidad de transacciones registradas en el día y momento de la recogida de taxi.

5.6. TaxiY

Este conjunto de datos contiene información sobre transacciones de recogida de taxis a combustión interna en un período específico. Cada registro incluye detalles como la fecha de recogida , el día de la semana

, el monto total de la transacciones y la cantidad de transacciones registradas.

- pickup_date: Fecha en la que se realizó la recogida de taxi.
- weekday: Día de la semana en el que tuvo lugar la recogida de taxi.
- total_amount: Monto total de las transacciones correspondiente a la fecha.
- row_count: Cantidad de transacciones registradas en el día y momento de la recogida de taxi.

5.7. **Calidad_del_aire**

Este conjunto de datos representa una amplia recopilación de información relacionada con varios aspectos ambientales y de salud en diferentes ubicaciones geográficas y a lo largo de distintos períodos de tiempo.

- Name: nombre de la medida que se está registrando.
- Measure: Indica la unidad de medida utilizada para registrar los datos.
- Measure Info: proporciona información adicional sobre la medida
- Geo Type Name: Representa el tipo de ubicación geográfica para la que se están registrando los datos.
- Geo Join ID: Identificador numérico asociado a la ubicación geográfica específica.
- Geo Place Name: El nombre del lugar geográfico, que puede ser un distrito municipal o un barrio específico.
- Time Period: El período de tiempo durante el cual se recopilaron los datos.
- Start_Date: La fecha de inicio del período de tiempo para el cual se están registrando los datos.
- Data Value: El valor numérico de la medida registrada en esa ubicación geográfica y período de tiempo específicos.

5.8. **NYCclima**

Este conjunto de datos representa información sobre estimaciones de tiempo, temperatura, precipitación y otras variables. Cada fila representa una hora específica y contiene varios valores correspondientes a esa hora.

- time: representa la fecha y la hora en la que se registraron los datos.
- hours: Representa la hora del día en formato "hora:minuto:segundo".
- temperature_2m (°C): Indica la temperatura a 2 metros sobre el suelo en grados Celsius en el momento específico.
- precipitation (mm): registra la cantidad de precipitación en milímetros en un momento específico.
- rain (mm): Registra la cantidad específica de lluvia en milímetros en ese momento.
- is_day (): indica si es de día o no en ese momento.

5.9. **taxis_tarifa**

Este dataset proporciona información detallada sobre la operación diaria de los taxis en términos de pasajeros, viajes, ingresos y métodos de pago.

- Fecha: Representa la fecha en la que se recopilaron los datos.
- Pasajeros por día: La cantidad de pasajeros transportados en ese día en particular.
- Viajes por día: El número total de viajes realizados en ese día.
- Tarifario por día: El total de tarifas acumuladas por los viajes realizados en ese día.
- Total recaudado por día: El monto total recaudado en ese día, incluyendo tanto pagos con tarjeta como en efectivo.
- Pago con tarjeta: La cantidad de dinero recaudado a través de pagos con tarjeta de crédito u otros métodos electrónicos.
- Pago con efectivo: La cantidad de dinero recaudado en efectivo.
- Tipo de Taxi: El tipo de taxi o servicio que se está analizando(Green, Yellow).

6. Análisis Exploratorio

En el proceso de realizar el Análisis Exploratorio de Datos (EDA), se ha evidenciado un notorio incremento en los niveles de contaminación sonora durante los últimos años. Específicamente, este fenómeno se muestra más acentuado en el área urbana de Nueva York. Este hallazgo sugiere una correlación significativa entre la contaminación sonora y los ruidos generados por los motores de los vehículos.
Dentro de las estadísticas descriptivas del volumen total de contaminación sonora en la ciudad de Nueva York, se destaca que el 50% de esta proviene directamente de los ruidos emitidos por los motores de los vehículos. Esta proporción resalta que exactamente la mitad de la contaminación sonora tiene su origen en los sonidos generados por los motores de los vehículos en circulación.

Además, se ha identificado una relación estrecha entre el tamaño de los motores de los vehículos y las emisiones de dióxido de carbono (CO2). Este patrón muestra una tendencia clara: a medida que el tamaño del motor aumenta, también lo hacen las emisiones de CO2, lo que resalta la importancia de abordar no solo la contaminación sonora, sino también las implicaciones ambientales asociadas con los vehículos de mayor envergadura.

Otro aspecto destacable del análisis es la concentración de actividad de taxis en las áreas residenciales. Se ha observado que las zonas con mayor densidad de movimiento de taxis coinciden con las áreas de los diferentes barrios de la ciudad. Esto sugiere un patrón de demanda de transporte en las zonas urbanas más concurridas, lo que podría contribuir de manera significativa a los niveles de contaminación sonora y a la complejidad del entorno acústico en estas áreas.

Este análisis exploratorio proporciona una visión inicial pero valiosa de la relación entre la contaminación sonora, los vehículos y los patrones de actividad en Manhattan. Estos resultados resaltan la importancia de abordar de manera integral la gestión del ruido ambiental y sus múltiples interacciones con el entorno urbano y los medios de transporte**.**

En cuanto al análisis de la calidad del aire, se centró en componentes clave como el Ozono (O3), Dióxido de Azufre (SO2), Dióxido de

Nitrógeno (NO_2) y Partículas Finas ($PM_{2.5}$). Estos componentes son fundamentales para evaluar el impacto en la salud y el medio ambiente debido a la contaminación del aire. Se realizó un análisis de los últimos 3 años recopilados, observando la evolución de los valores de Ozono (O_3) que se mantienen relativamente constantes en el tiempo. También se notaron picos de Dióxido de Nitrógeno (NO_2) en estaciones específicas del año y una leve reducción en los valores de Partículas Finas ($PM_{2.5}$) a lo largo del tiempo.

En relación con el tráfico vehicular, se analizaron componentes clave, destacando la diferencia en la densidad de tráfico entre vehículos de pasajeros y camiones. La circulación de vehículos de pasajeros es notablemente mayor en comparación con la de camiones, lo cual refleja una dinámica importante en la movilidad urbana.

Se procedió a realizar visualizaciones por diferentes tipos de áreas geográficas en la ciudad de Nueva York, como Distritos Comunitarios (CD), Boroughs, Uniformed Hospital Fund Area (UHF) y Citywide. Estas visualizaciones indicaron que las áreas más pequeñas (CD) concentran más los componentes clave de la calidad del aire.

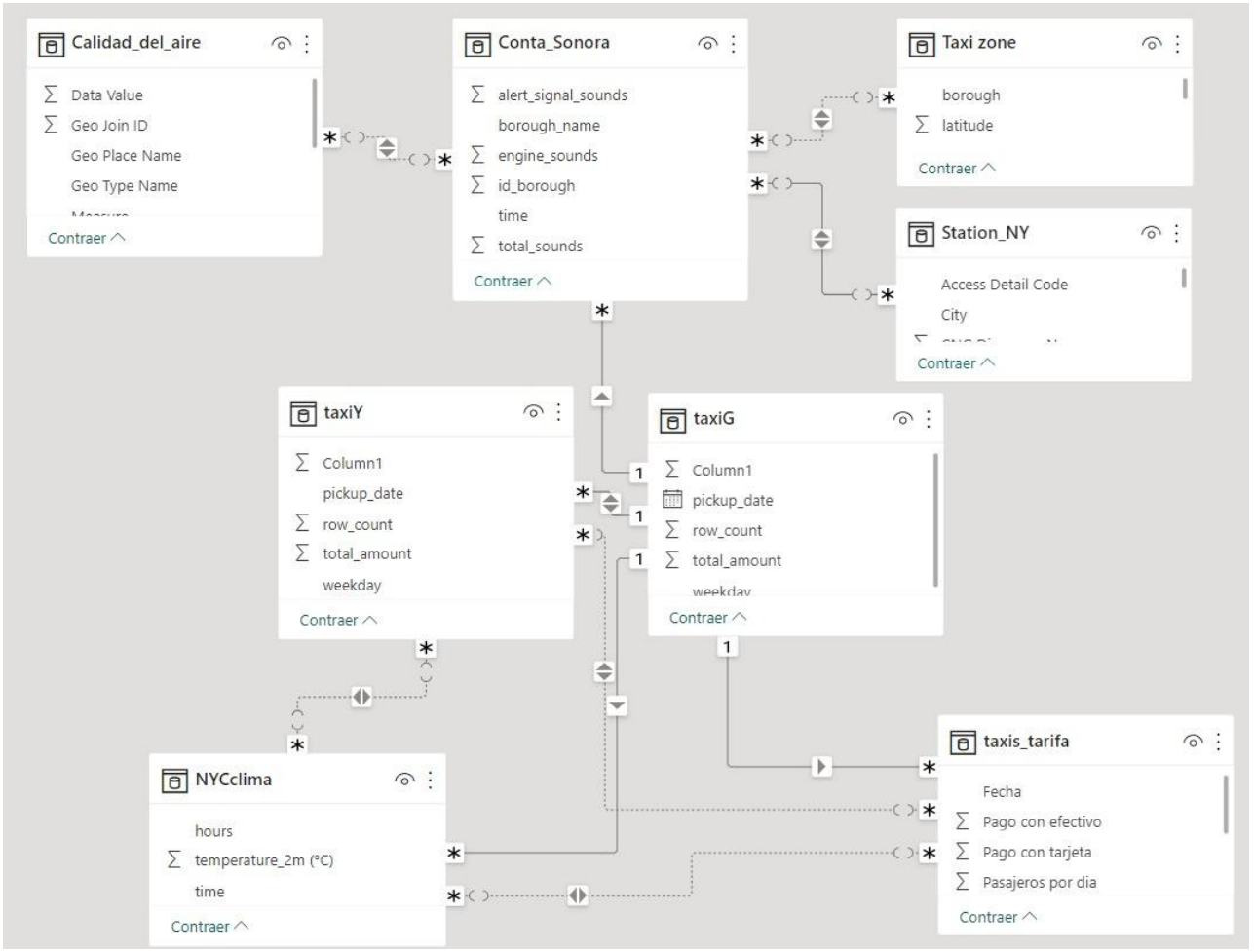
En cuanto al transporte, se observó que los autos de combustión son los más utilizados, aunque se notó una incipiente implementación de autos verdes. Esta transición hacia vehículos más sostenibles es reciente y se refleja en los datos.

Se exploró la correlación entre la cantidad de viajes de taxis y las precipitaciones. En el caso de taxis verdes, no se observaron aumentos notables durante precipitaciones elevadas, mientras que en taxis amarillos hubo un pequeño incremento durante periodos de precipitaciones más altas. Además, se planteó la posibilidad de que la temperatura tenga una influencia más fuerte en la cantidad de viajes de taxis que las precipitaciones, basado en las observaciones de los gráficos.

En resumen, el Análisis Exploratorio de Datos proporcionó una comprensión más profunda de cómo los componentes clave, la densidad de tráfico y las condiciones climáticas influyen en diversos aspectos de la calidad del aire y el transporte en la ciudad de Nueva

York. Aunque algunas comparaciones y análisis específicos no pudieron realizarse debido a la falta de datos, se extrajeron valiosas conclusiones sobre la dinámica urbana y la interacción entre estos factores.

7. Creación del Data Warehouse (Diagrama Entidad Relación)



8. Tareas por realizar

8.1 Automatizar el Data Warehouse

Para lograr la automatización del data warehouse, implementaremos el uso de Airflow. En el cuaderno, configuraremos un flujo de trabajo que facilitará la ejecución de nuestro proceso ETL, asegurando así la transformación y carga eficiente de los datos depurados en MySQL.

8.2 Carga incremental del Data Warehouse

Airflow realiza una carga incremental al emplear un enfoque inteligente para identificar y procesar solo los nuevos datos o los que han cambiado desde la última ejecución. Utilizando marcas de tiempo y metadatos, Airflow compara los registros existentes con los nuevos datos, permitiendo una actualización eficiente de la base de datos sin tener que procesar nuevamente todo el conjunto de datos. Esto agiliza el proceso y optimiza el uso de recursos al reducir la carga de trabajo y el tiempo requerido para mantener la integridad y actualidad de la información en el sistema.