Diccionario de datos

Proyecto **EcoTaxiNYC**
09 de abril de 2024
Versión **1.0**

Sobre este documento

El presente diccionario de base de datos tiene como objetivo documentar la estructura y el contenido de las bases de datos que albergan información sobre los viajes en taxi, el clima, la contaminación sonora y la contaminación por CO2 en la ciudad de Nueva York (NYC). Esta información es fundamental para comprender los patrones de movilidad urbana, el impacto ambiental del transporte de motor a combustión interna en comparación con el motor eléctrico y la calidad del aire en la ciudad.

Este diccionario está dirigido a administradores de bases de datos, analistas de datos, científicos de datos y cualquier persona que necesite comprender el significado y la organización de los datos contenidos en las bases de datos.

Esperamos que este diccionario de base de datos facilite el acceso y la comprensión de la información vital sobre la ciudad de Nueva York. La información contenida en este documento es un recurso valioso para la toma de decisiones informadas sobre transporte, medio ambiente y clima en la ciudad.

Descripción de las bases de datos

La base de datos implementada en BigQuery de Google Cloud Platform (GCP) tiene como objetivo centralizar y analizar información relacionada con el proyecto de creación de una flota de taxis eléctricos en NYC.

Fuentes de datos

Los datos se obtienen de diversas fuentes, incluyendo:

- **Datos de viajes en taxi**: Estos datos provienen de la Comisión de Taxis y Limusinas de la Ciudad de Nueva York (NYC Taxi & Limousine Commission) y contienen información sobre cada viaje en taxi, como la fecha, hora, origen, destino, distancia recorrida y tarifa.
- **Datos climáticos**: Los datos climáticos se obtienen de la Administración Nacional Oceánica y Atmosférica (NOAA) y del Servicio Meteorológico Nacional (NWS) e incluyen información sobre temperatura, humedad, precipitación, presión atmosférica y velocidad del viento.

- **Datos de contaminación sonora**: Los datos de la contaminación sonora se obtienen del Departamento de Protección Ambiental de la Ciudad de Nueva York (NYC Department of Environmental Protection) y contienen información sobre los niveles de ruido en diferentes puntos de la ciudad.
- **Datos de contaminación por CO2**: Los datos de la contaminación por CO2 se obtienen de la Agencia de Protección Ambiental de los Estados Unidos (EPA) y del Proyecto de Carbón Global (Global Carbon Project) e incluyen información sobre las emisiones de CO2 en la ciudad.

Principales características

- **Base de datos no relacional**. Se utiliza BigQuery, un servicio de base de datos no relacional de Google Cloud Platform, para almacenar y gestionar los datos, es decir, no sigue el modelo relacional propuesto originalmente por E.F. Codd. BigQuery permite almacenar datos de forma no estructurada en tablas, similar a una base de datos multidimensional.
- **Tablas**. La base de datos se compone de las siguientes tablas:
    - **Noise**: Contiene datos sobre la contaminación sonora en la ciudad.
    - **CarbonEmissionNYC**: Contiene datos sobre las emisiones de CO2 en la ciudad.
    - **Borough**: Contiene información sobre los distritos de la ciudad.
    - **ElectricCarSpec**: Contiene especificaciones de los vehículos eléctricos.
    - **VehicleAndDriverReport**: Contiene información sobre los vehículos y conductores.
    - **TaxiZone**: Contiene información sobre las zonas de taxis en la ciudad.
    - **AirQuality**: Contiene datos sobre la calidad del aire en la ciudad.
    - **Trip**: Contiene información sobre cada viaje en taxi.
    - **Weather**: Contiene datos climáticos históricos de la ciudad.
- **Análisis de datos**. Se utilizan herramientas de análisis EDA para explorar y comprender los datos, identificar patrones y tendencias, y tomar decisiones informadas.

Esquema de las tablas

Se detalla la estructura de cada tabla en las bases de datos, incluyendo los nombres de las columnas, los tipos de datos y las relaciones entre las tablas.

Tabla Noise

Tabla: **Noise** Grupo: **data_clean**		Descripción: Contiene datos sobre la contaminación sonora en la ciudad			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	borough	INTEGER	NULLABLE	-	Distrito de NYC
	block	INTEGER	NULLABLE	-	El bloque de NYC en el que se encuentra el sensor. Dígitos del 2 al 6 en el sistema de números de parcelas de NYC de 10 dígitos conocido como BBL
	latitude	FLOAT	NULLABLE	-	Latitud
	longitude	FLOAT	NULLABLE	-	Longitud
	year	INTEGER	NULLABLE	-	Año en el que ocurrió el sonido
	week	INTEGER	NULLABLE	-	Semana en la que ocurrió el sonido
	day	INTEGER	NULLABLE	-	Día en el que ocurrió el sonido
	hour	INTEGER	NULLABLE	-	Hora en la que ocurrió el sonido
	smallSoundingEngine	INTEGER	NULLABLE	-	Presencia de sonido de motor pequeño
	mediumSoundingEngine	INTEGER	NULLABLE	-	Presencia de sonido de motor mediano

Extracción (Cloud Storage/archivo)	Transformación (Cloud function)	Carga (BigQuery)
data_ruidos_new_york/SONYCUrbanSoundTagging.csv	process_noise (script Python)	Noise (Tabla)

Tabla AirQuality

Tabla: **AirQuality** Grupo: **data_clean**		Descripción: Contiene datos sobre la calidad del aire en la ciudad			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	year	INTEGER	NULLABLE	-	Año en el que se capturó el dato
	borough	INTEGER	NULLABLE	-	Distrito de NYC
	fineParticlesPM25	FLOAT	NULLABLE	-	Concentración media de partículas de 2.5 microns, en mcg/m3
	nitrogenDioxideNO2	FLOAT	NULLABLE	-	Concentración media de NO2 en ppb
	ozoneO3	FLOAT	NULLABLE	-	Concentración media de ozono en ppb

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_calidad_aire_new_york/Air_Quality.csv	process_air_quality	AirQuality

Tabla Borough

Tabla: **Borough** Grupo: **data_clean**		Descripción: Refleja la división político/territorial en distritos (boroughs) de la ciudad de Nueva York			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	id	INTEGER	NULLABLE	-	Identificador único del distrito de Nueva York (NYC)
	name	INTEGER	NULLABLE	-	Nombre del distrito de NYC
	county	FLOAT	NULLABLE	-	Condado al que pertenece el distrito

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
N/A	N/A	Borough

Tabla CarbonEmissionNYC

Tabla: **CarbonEmissionNYC** Grupo: **data_clean**		Descripción**:** Refleja las emisiones anuales de CO2 de la ciudad de Nueva York, según sector, categoría y fuente.			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	Sectors Sector	STRING	NULLABLE	-	Sector de actividad de emisiones
	Category Full	STRING	NULLABLE	-	Categoría de actividad
	Source Label	STRING	NULLABLE	-	Subcategoría de actividad
	Año	INTEGER	NULLABLE	-	Año correspondiente
	Consumed	FLOAT	NULLABLE	-	Toneladas de CO2 emitidas

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_co2_new_york/CY_2005_CY_2022_citywide.xlsx	process_emission	CarbonEmissionNYC

Tabla TaxiZone

Tabla: **TaxiZone** Grupo: **data_clean**		Descripción**:** Refleja las zonas de taxis de la ciudad de Nueva York.			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	LocationID	INTEGER	NOTNULL	-	ID de la zona
	borough	STRING	NULLABLE	-	Distrito de NYC al que pertenece la zona
	zone	STRING	NULLABLE	-	Zona de taxis
	longitud	FLOAT	NULLABLE	-	Longitud geoespacial del centroide
	latitud	FLOAT	NULLABLE	-	Latitud geoespacial del centroide
	geometry	POLYGON	NULLABLE	-	Polígono que describe la forma geoespacial de la zona de taxis

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_taxi_zone/taxi_zones.json	process_taxi_zone	TaxiZone

Tabla Trip

Tabla: **Trip** Grupo: **data_clean**		Descripción**:** Refleja los viajes realizados en las zonas de taxis de la ciudad de Nueva York.			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	year	INT	NULLABLE	-	Año de los viajes
	month	INT	NULLABLE	-	Mes de los viajes
	day	INT	NULLABLE	-	Día del mes
	dayOfWeek	INT	NULLABLE	-	Día de la semana
	hour	INT	NULLABLE	-	Franja horaria
	puLocationID	INT	NULLABLE	-	ID de zona de taxi donde comienza el viaje
	doLocationID	INT	NULLABLE	-	ID de zona de taxi donde finaliza el viaje
	timeOut	INT	NULLABLE	-	Tiempo de demora entre la solicitud del viaje y la llegada del vehículo.
	travelTime	INT	NULLABLE	-	Tiempo de viaje en segundos
	serviceNumber	INT	NULLABLE	-	Cantidad de servicios en el rango horario
	tripMiles	FLOAT	NULLABLE	-	Millas del recorrido
	fareSurcharges	FLOAT	NULLABLE	-	Suma de recargos a la tarifa base
	baseFare	FLOAT	NULLABLE	-	Suma de la tarifa base y el impuesto “sales fare”

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_taxis_amarillos data_taxis_verdes data_alquiler_gran_volumen	process_trip	Trip

Tabla Weather

Tabla: **Weather** Grupo: **data_clean**		Descripción**:** Datos climáticos históricos de NYC			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	fecha	DATETIME	NULLABLE	-	Fecha del registro
	hora	INT	NULLABLE	-	Franja horaria del registro
	temperatura	FLOAT	NULLABLE	-	Temperatura en °C a 2 metros de altura
	humedad	FLOAT	NULLABLE	-	Humedad relativa a 2 metros de altura
	lluvia	FLOAT	NULLABLE	-	Cantidad de lluvia caída la hora anterior, en mm
	nieve	FLOAT	NULLABLE	-	Cantidad de nieve caída la hora anterior, en cm

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_climatico	process_weather	Weather

Tabla ElectricCarSpec

Tabla: **ElectricCarSpec** Grupo: **data_clean**		Descripción**:** Datos sobre características de vehículos eléctricos			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	carModel	STRING	NULLABLE		Modelo del vehículo
	timeFrom0To100	FLOAT	NULLABLE		Aceleración de 0 a 100 km/h en segundos
	maxSpeed	INT	NULLABLE		Velocidad máxima en Km/h
	range	INT	NULLABLE		Rango en km
	fastCharging	INT	NULLABLE		Carga en km/h
	price	INT	NULLABLE		Precio en euros

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_electric_car	process_electric_car	ElectricCarSpec

Tabla AggregatedReport

Tabla: **Aggregate Report** Grupo: **data clean**		Descripción**: Contiene información de conductores y vehículos**			
Clave	Nombre del campo	Tipo de dato	Valor nulo	Único	Descripción
	year	INTEGER	NULLABLE	-	Año de los servicios agregados
	month	INTEGER	NULLABLE	-	Mes de los servicios agregados
	licenseClass	STRING	NULLABLE	-	Tipo de Vehículo (Yellow, Green, FHV y High Volume (Uber y Lyft))
	uniqueDrivers	FLOAT	NULLABLE	-	El número total de conductores únicos que registraron un viaje cada mes (promedio mes)
	uniqueVehicles	FLOAT	NULLABLE	-	El número de vehículos únicos que registraron al menos un viaje en el mes (promedio mes)
	vehiclesPerDay	FLOAT	NULLABLE	-	El número de vehículos únicos que registraron al menos un viaje al día (promedio día)
	avgDaysVehiclesOnRoad	FLOAT	NULLABLE	-	Número promedio de días que el vehículo pasó en la carretera x mes (días de operación)
	avgHoursPerDayPerVehicle	FLOAT	NULLABLE	-	El promedio de horas en las que un vehículo registró un viaje (horas de operación día)
	avgDaysDriversOnRoad	FLOAT	NULLABLE	-	El número promedio de días que cada conductor registró un viaje (días de trabajo promedio)
	avgHoursPerDayPerDriver	FLOAT	NULLABLE	-	El número promedio de horas por cada conductor (horas trabajadas promedio por día)
	avgMinutesPerTrip	FLOAT	NULLABLE	-	Tiempo promedio de viaje (minutos) desde el medidor encendido hasta apagado
	percentOfTripsPaidWithCreditCard	FLOAT	NULLABLE	-	Viajes en los que el pasajero pagó con tarjeta de crédito del número total de viajes
	tripsPerDayShared	FLOAT	NULLABLE	-	Número medio de viajes compartidos registrados cada día

Extracción (Cloud Storage / file)	Transformación (Python Cloud Function)	Carga (Tabla BigQuery)
data_reports_monthly.csv	process_aggregated_report	AggregatedReport

Glosario de términos

Se define la terminología específica utilizada en las bases de datos, asegurando una comprensión clara y consistente de los datos.

C

CO2: Dióxido de carbono. Gas procedente de la combustión de combustibles fósiles. Pág. 2, 3, 6

N

NO2: Dióxido de nitrógeno. Gas procedente de la combustión de combustibles fósiles. Pág 5

O

O3: Ozono Gas generado en la tropósfera por los gases de la combustión. Pág. 5

P

PM2.5: Partículas liberadas durante la combustión incompleta de combustibles fósiles.
ppb: partículas por billón. Pág 5

T

Tarifa base: La tarifa base es la tarifa inicial (en NYC es de $3.00) que se cobra al iniciar un viaje en taxi. Se le suman varios recargos. Pág 1, 7

Información de contacto

Se proporcionan los datos de contacto del administrador de la base de datos para cualquier consulta o aclaración sobre la información contenida en el diccionario.

Alter Caimi ✉ caimialter@gmail.com
Carlos Masea ✉ macea42@gmail.com
Luis Rojas ✉ ldavidrd@gmail.com
Rafael Balestrini ✉ rafaelbalestrini@gmail.com

ÍNDICE

Los datos que no cuentan con un diccionario corren el riesgo de ser erróneamente interpretados y utilizados. En algunos casos los datos son inutilizables, ya que su interpretación se vuelve imposible.