**Procesamiento de Datos**

**Analista de datos:** Luis Fernando Martínez

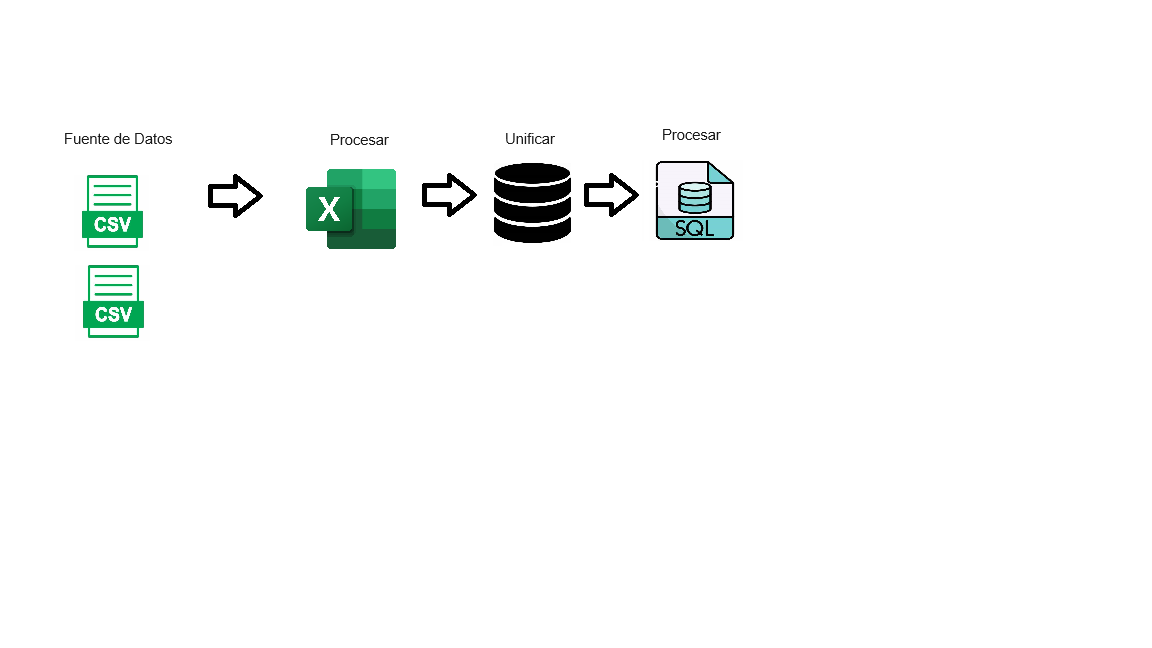
**Fecha de creación:** 2024/07/21

**Fecha de actualización:** 2025/01/04

**Herramientas**

Para determinar las herramientas a utilizar para el procesamiento de datos se debe tener en cuenta el estado de las fuentes de datos, formato, cantidad de registros y cantidad de variables. La fuente de datos se divide en 12 archivos fuente, en formato csv, en su totaliad contienen 5743290 registros, y cada uno se compone de 13 variables.

Debido a la organización de los datos (12 archivos independientes) que no superan 1 millón de registros cada uno, se procesarán individualmente utilizando Excel (limpieza), se verificará el procesamiento y se unificarán en un solo archivo o se unificarán en una tabla en un SGBD. Debido a la cantidad total de registros totales, se usará SQL para un procesamiento posterior con todos los datos unificados.



Para aplicar los temas vistos en el curso se utilizará Big Query como SGBD en su versión gratuita.

**Limpieza de datos**

Debido a la gran cantidad de registros por archivo, en el preprocesamiento realizado en Excel solo se eliminarán registros que presenten atributos en blanco y que no se puedan completar con información propia del dataset, siempre y cuando estos atributos sean vitales para el análisis.

Posteriormente se subirán los archivos a Big Query para el procesamiento final de los datos.

Por las limitaciones de Big Query gratuito, las tablas generadas mediante archivos no pueden superar los 100 MB, afortunadamente para este caso, después del preprocesamiento, cada archivo tendrá un tamaño menor al del umbral, en caso de tener archivos con tamaño mayor a 100 MB sería necesario usar Google Cloud Storage para subir en un bucket todos los datos y luego ser utilizados en Big Query.

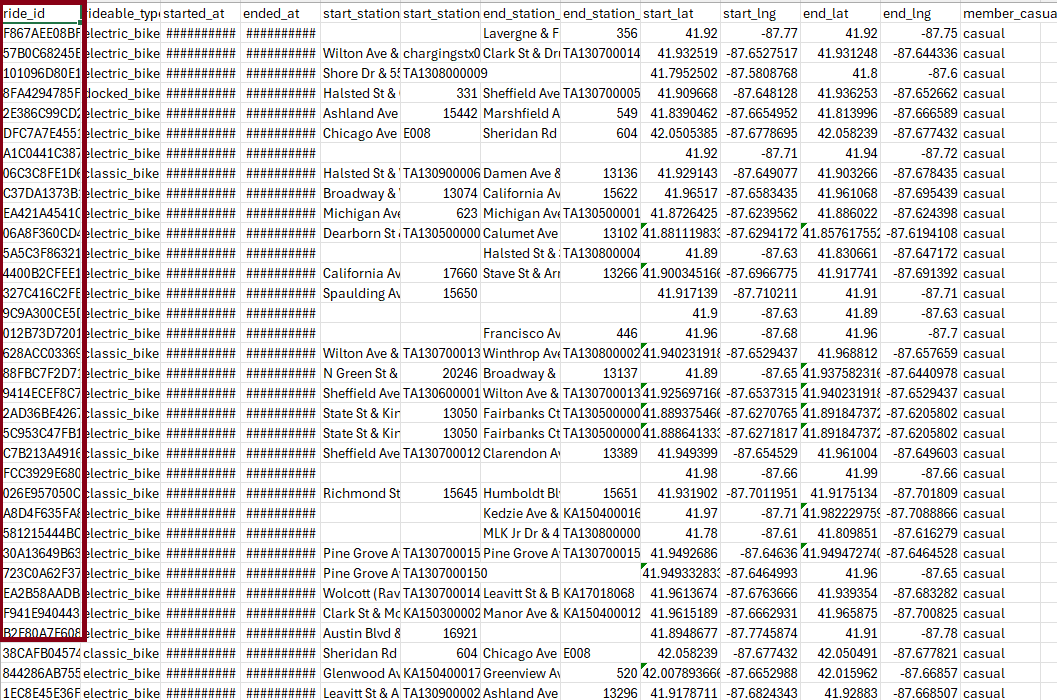
Eliminación de registros con atributos en blanco

Atributos vitales para el análisis:

* rideable\_type
* started\_at
* ended\_at
* start\_station\_name
* start\_station\_id
* end\_station\_name
* end\_station\_id
* start\_lat
* start\_lng
* end\_lat
* end\_lng
* member\_casual

Los demás atributos serán eliminados, siendo estos:

* ride\_id: este atributo no aporta información valiosa ya que es la identificación individual del evento “viaje”, sería valioso si fuera la identificación del usuario, pero por protección de cada usuario esta información no está disponible ni se puede buscar o inferir.



Los registros que serán eliminados deben cumplir alguna de las siguientes reglas:

* Registro en blanco: No existe ningún dato
* Alguno de los atributos vitales para el análisis está en blanco y no se puede inferir desde el mismo dataset.

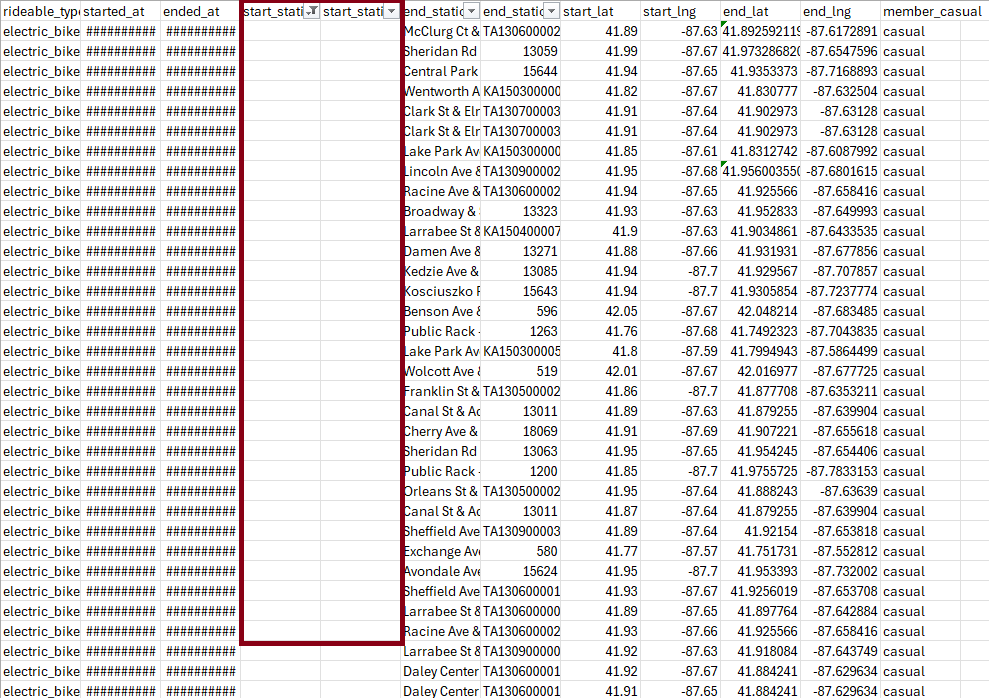
Siguiendo las reglas mencionadas se eliminaron los registros en blanco de cada archivo, a continuación, mediante un filtrado por columnas, se buscaron registros que tuvieran alguno de los atributos vitales para el análisis en blanco.

Al finalizar la exploración de atributos en blanco, 6 campos presentaban casos de registros sin información acerca de estos, los cuales fueron:

* start\_station\_name
* start\_station\_id
* end\_station\_name
* end\_station\_id
* end\_lat
* end\_lng

Puesto que tanto el nombre, id, latitud y longitud se pueden inferir del mismo dataset, solo se eliminarán los registros que cumplan con alguna de las siguientes subreglas:

* El registro no tiene “start\_station\_name” y no tiene “start\_station\_id”.
* El registro no tiene “end\_station\_name” y no tiene “end\_station\_id”.

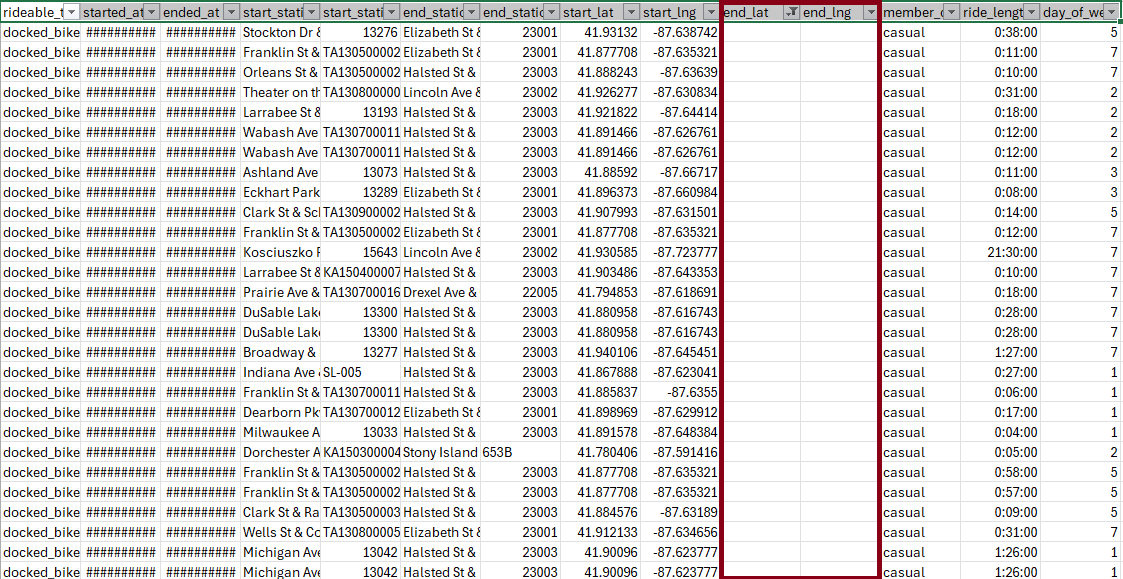


**Tabla

Descripción generada automáticamente**

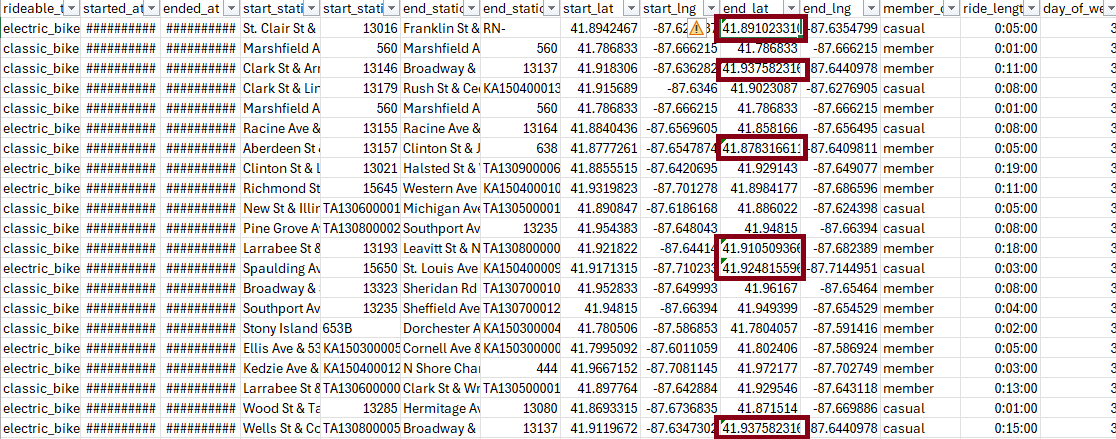
**Verificación de datos**

Mediante el filtrado de datos en Excel, se buscaron registros que aún tuvieran alguno de los campos “start\_station\_name”, “start\_station\_id”, “end\_station\_name”, o “end\_station\_id” en blanco, en este caso no quedaban registros con valores en blanco, por lo que se busca un método para completar los registros que aún tienen los campos “end\_lat” o “end\_lng” en blanco, con los datos propios del dataset.



Haciendo uso de la función “BUSCARV” (VLOOKUP) se completan los valores faltantes para “end\_lat” y “end\_lng”.

Al terminar la tarea anterior, se corrige el tipo de datos de algunas latitudes.



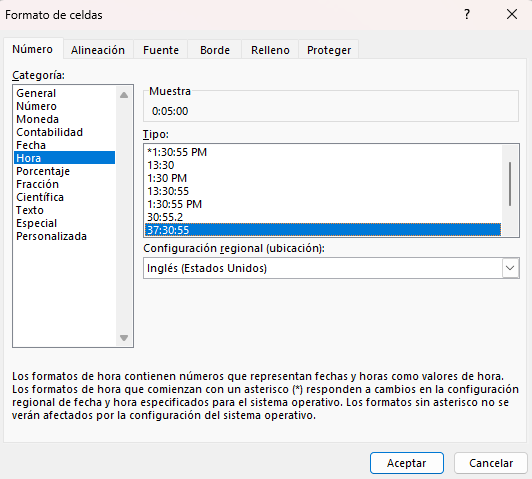
**Adición de columnas**

Para hacer el análisis completo y de manera más rápida, se agregarán un par de columnas al dataset, las cuales serán:

* ride\_length: Muestra el tiempo del viaje en formato de hora HH:MM:SS.
* day\_of\_week: Muestra el día de la semana en el que se realizó el viaje.

Adición columna “ride\_length”

Esta columna muestra el tiempo transcurrido durante el viaje, y se calcula restando la columna “started\_at” de la columna “ended\_at”, luego se aplica formato de hora HH:MM:SS (37:30:55).



Tabla

Descripción generada automáticamente

Eliminación de registros cuyo tiempo de llegada sea menor que el tiempo de salida.

Al verificar los valores obtenidos en “ride\_length”, se hallan registros con valores negativos, lo que sugiere una inconsistencia en las fechas de llegada y salida del registro, estos registros deben ser eliminados.

Atributos por verificar:

* started\_at
* ended\_at

Los registros que serán eliminados deben cumplir la siguiente condición:

* Hora de llegada es menor que la hora de salida: ended\_at < started\_at

Para encontrar los registros que cumplen la condición y poder filtrarlos, se utiliza una columna auxiliar “started\_greater\_than\_ended” que tendrá uno de dos valores, “TRUE” si cumple la condición o “FALSE” si no la cumple, usando la función IF(ended\_at < started\_at) en cada registro. Al eliminar los registros se eliminará la columna auxiliar.

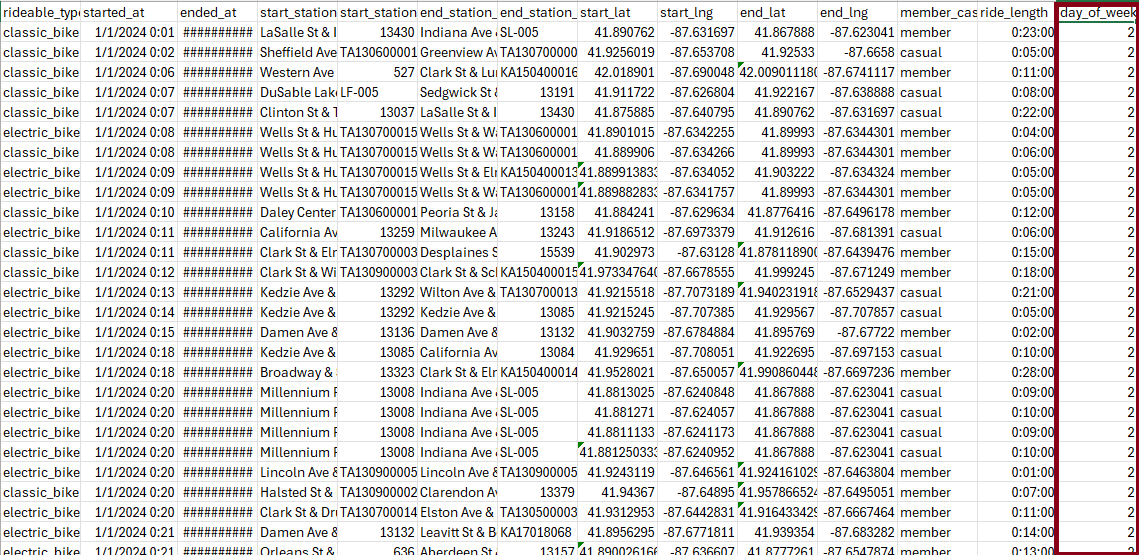
Tabla

Descripción generada automáticamente con confianza media



Adición columna “day\_of\_week”

Esta columna muestra el día de la semana en la que se realizó el viaje, donde domingo:1 y sabado:7, se utiliza la función “diasem” sobre el campo “started\_at” y en formato de celda “general”.



Calendario

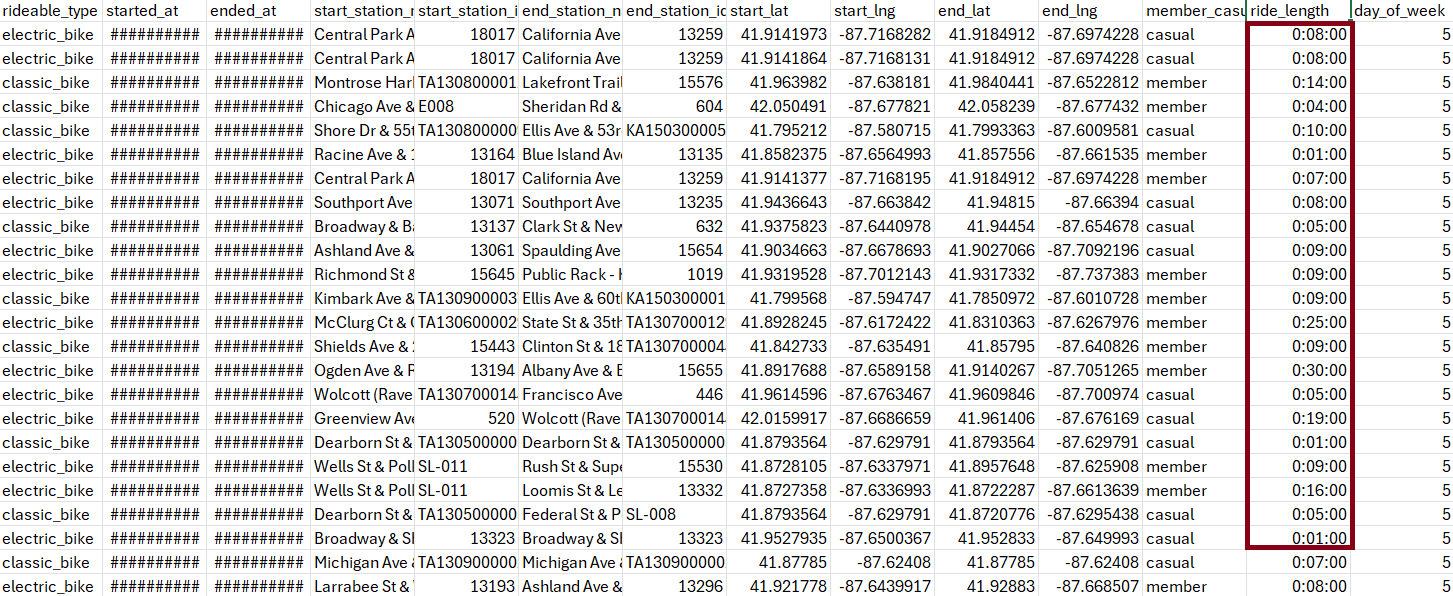
Descripción generada automáticamente

Como se evidencia, en el formato especificado, el número 2 corresponde al lunes, en la imagen se ven datos de 1/1/2024 el cual corresponde a un lunes.

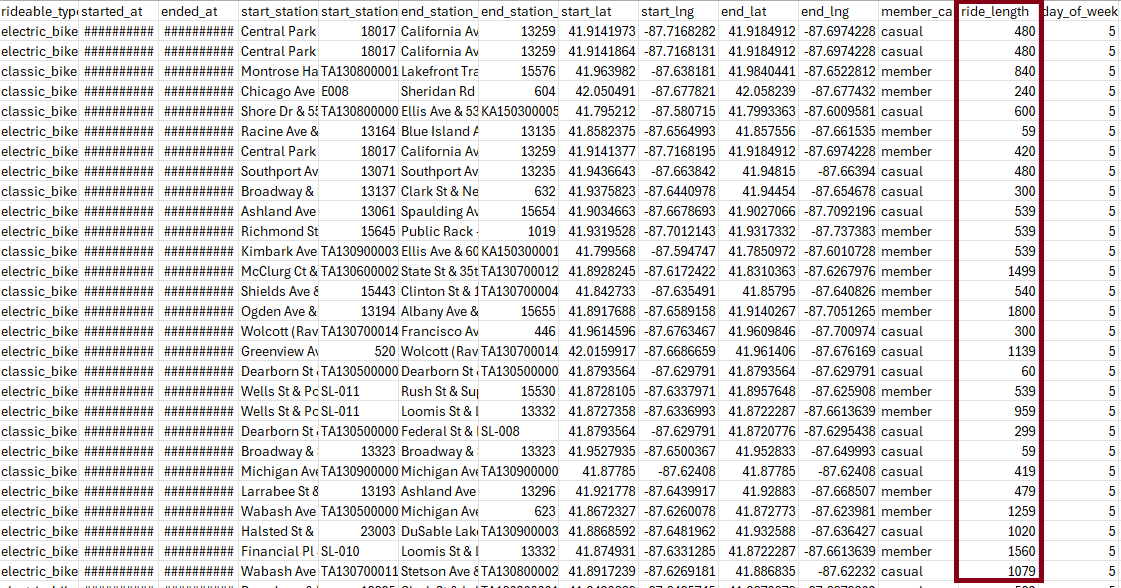
**Carga de datos procesados a BigQuery**

Cada archivo .csv se subirá a BigQuery como una tabla independiente, con el nombre del mes y el año. BigQuery tiene un límite de tamaño de archivo de 100 MB, y ya que existe por lo menos un archivo que sobrepasa este límite, se decide modificar el formato de la columna “ride\_length” para mostrar solo los segundos transcurridos en el viaje, lo anterior permitirá reducir el tamaño del documento.

Para lograr el formato requerido se utiliza la fórmula:



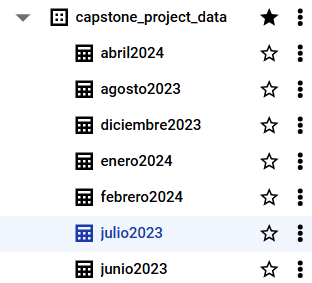
Se cambia al formato de solo segundos



Al final del preprocesamiento de datos utilizando Excel se tienen los siguientes datos:

|  |  |  |
| --- | --- | --- |
| **Nombre Archivo** | **Tamaño [MB]** | **Registros** |
| 202306-divvy-tripdata.csv | 90,0 | 534768 |
| 202307-divvy-tripdata.csv | 96,5 | 574023 |
| 202308-divvy-tripdata.csv | 98,2 | 584957 |
| 202309-divvy-tripdata.csv | 85,0 | 506634 |
| 202310-divvy-tripdata.csv | 68,4 | 403781 |
| 202311-divvy-tripdata.csv | 46,5 | 274768 |
| 202312-divvy-tripdata.csv | 28,3 | 167142 |
| 202401-divvy-tripdata.csv | 19,1 | 113808 |
| 202402-divvy-tripdata.csv | 30,9 | 184736 |
| 202403-divvy-tripdata.csv | 38,6 | 230278 |
| 202404-divvy-tripdata.csv | 50,0 | 297798 |
| 202405-divvy-tripdata.csv | 74,2 | 442168 |
| Total |  | 4314861 |

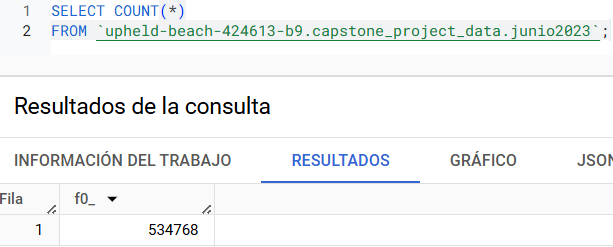
Al finalizar la transformación de la columna “ride\_length” se cargan los archivos a un dataset de BigQuery llamado “capstone\_project\_data”, donde cada archivo se sube como una tabla independiente con nomenclatura “mesaño”.



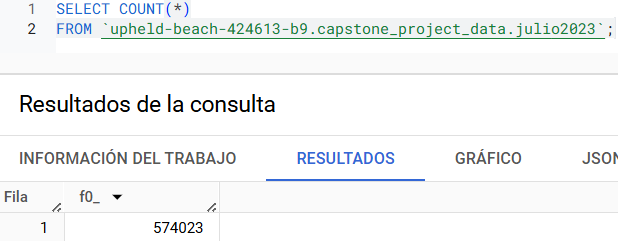
**Verificación de cantidad de registros cargados a BigQuery**

|  |  |  |
| --- | --- | --- |
| **Nombre Archivo o tabla** | **Registros csv** | **Registros BigQuery** |
| 202306-divvy-tripdata.csv | 534768 | 534768 |
| 202307-divvy-tripdata.csv | 574023 | 574023 |
| 202308-divvy-tripdata.csv | 584957 | 584957 |
| 202309-divvy-tripdata.csv | 506634 | 506634 |
| 202310-divvy-tripdata.csv | 403781 | 403781 |
| 202311-divvy-tripdata.csv | 274768 | 274768 |
| 202312-divvy-tripdata.csv | 167142 | 167142 |
| 202401-divvy-tripdata.csv | 113808 | 113808 |
| 202402-divvy-tripdata.csv | 184736 | 184736 |
| 202403-divvy-tripdata.csv | 230278 | 230278 |
| 202404-divvy-tripdata.csv | 297798 | 297798 |
| 202405-divvy-tripdata.csv | 442168 | 442168 |
| Total | 4314861 | 4314861 |

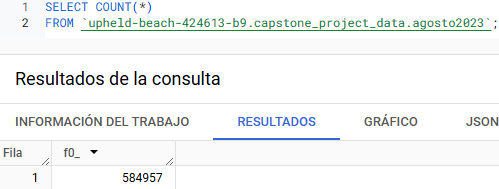
Cantidad de registros junio de 2023 en BigQuery



Cantidad de registros julio de 2023 en BigQuery



Cantidad de registros agosto de 2023 en BigQuery



Cantidad de registros septiembre de 2023 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

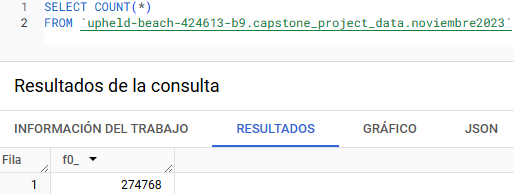
Descripción generada automáticamente

Cantidad de registros octubre de 2023 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros noviembre de 2023 en BigQuery



Cantidad de registros diciembre de 2023 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros enero de 2024 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros febrero de 2024 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros marzo de 2024 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros abril de 2024 en BigQuery

Interfaz de usuario gráfica, Aplicación

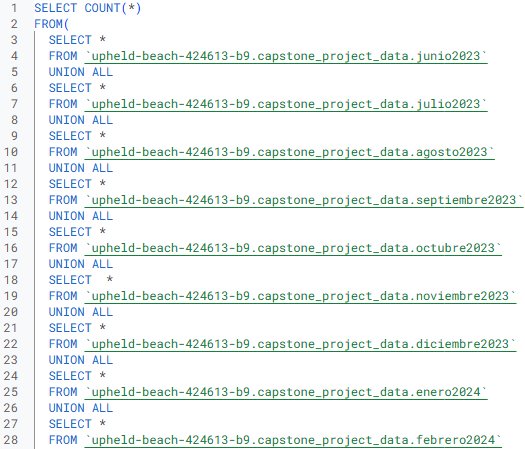
Descripción generada automáticamente

Cantidad de registros mayo de 2024 en BigQuery

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

Cantidad de registros consolidados en BigQuery sin remover duplicados

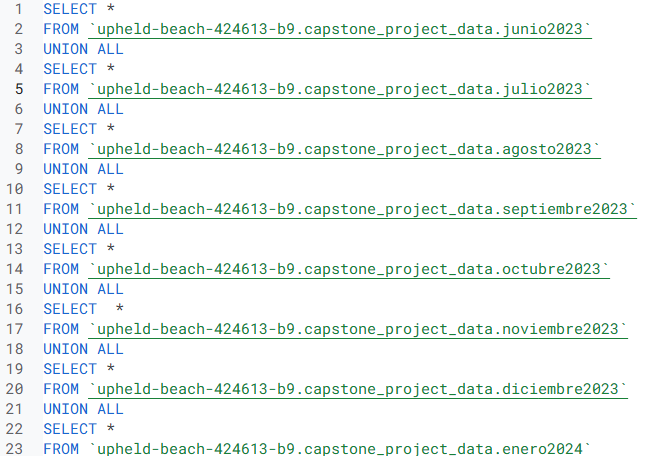


Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Consolidación de datos en BigQuery**

Se unirán los datos de cada tabla independiente del dataset “capstone\_project\_data” de BigQuery en una sola tabla utilizando SQL, la tabla se llamará “consolidado\_junio2023\_mayo2024\_raw”. Como las limitaciones de la versión gratis de BigQuery no permiten acciones de escritura, se guardará la consulta como una tabla nueva.



Captura de pantalla de un celular

Descripción generada automáticamente

Imagen que contiene Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

**Remoción de registros duplicados**

No existe un criterio oficial para determinar cuándo un registro se debe considerar duplicado. Para este proyecto se buscó aplicando formato condicional de Excel a la columna “ride\_id”, se observó que no existía ningún valor repetido para dos registros diferentes, y al no existir un requerimiento específico, no se eliminará ningún registro, aunque al omitir la anterior columna para hacer una comparación sí resultarían registros duplicados, podría explicarse si dos personas toman bicicletas al mismo tiempo, en la misma estación, y llegan al mismo tiempo a la misma estación.

**Estandarización de campos de texto**

Las columnas de tipo texto serán estandarizadas al remover espacios en blanco extra y/o al transformar los textos a solo minúsculas, debido a las limitaciones de BigQuery anteriormente mencionadas, la transformación se realizará con un SELECT y guardando los resultados en una tabla nueva llamada “consolidado\_junio2023\_mayo2024\_transformed”.

Columnas a las cuales se les removerán espacios en blanco extra:

* rideable\_type
* start\_station\_name
* start\_station\_id
* end\_station\_name
* end\_station\_id
* member\_casual

Columnas que serán transformadas a solo minúsculas:

* rideable\_type
* start\_station\_name
* end\_station\_name
* member\_casual

Texto

Descripción generada automáticamente

