PROYECTO DE INTEGRACIÓN Y ANÁLISIS DE DATOS GEOGRÁFICOS USANDO SSIS

ING. Silva Parraguez Maximo

I. ENTENDIENDO LOS DATOS Y EL PROYECTO.

1. Fuente de datos:

Para este proyecto, se necesitará datos Geográficos, como Shapefiles (archivos que almacenan datos geográficos) y datos de GPS.

- Shapefiles: me serán útiles para representar rutas de transporte, límites geográficos, etc.
- Datos de GPS: Me servirán para analizar movimientos o ubicaciones específicas.

Para llevar a cabo este proyecto, he recopilado dos tipos principales de datos geográficos relacionados con la Ciudad de México: los archivos Shapefile y los datos de GPS.

- **Shapefiles**:

  Los Shapefiles son archivos que contienen información geográfica vectorial, como líneas, puntos y polígonos, que representan elementos geográficos. En este proyecto, utilizaré los Shapefiles para delinear rutas de transporte y límites geográficos dentro de la Ciudad de México. Estos archivos los obtuve de **Geofabrik**, una empresa que ofrece datos de **OpenStreetMap** en diferentes formatos. Específicamente, descargué los Shapefiles desde su servidor de descargas para México https://download.geofabrik.de/north-america.html. Estos archivos proporcionan una representación detallada de la infraestructura vial, áreas urbanas y otras características geográficas relevantes para el análisis.

- **Datos de GPS**:

  Para analizar movimientos y ubicaciones específicas, recurrí a un conjunto de datos de rutas de taxis en la Ciudad de México. Este dataset, disponible en Kaggle https://www.kaggle.com/datasets/mnavas/taxi-routes-for-mexico-city-and-quito?select=mex_clean.csv, fue recopilado mediante la aplicación EC Taximeter entre junio del 2016 y julio del 2017 y contiene información detallada sobre trayectos de taxis, incluyendo coordenadas GPS, tiempos de viaje y distancias recorridas. Estos datos me permitirán estudiar patrones de movilidad y comportamiento del transporte en la ciudad, aportando una perspectiva práctica y actualizada al análisis geoespacial del proyecto.

2. Objetivo:

El objetivo de este proyecto es desarrollar un proceso ETL en SSIS para integrar, transformar y analizar datos geográficos relacionados con rutas de transporte público, con el fin de generar insights útiles para el análisis de rutas, su optimización y la toma de decisiones.

II. DESARROLLO DEL PROYECTO

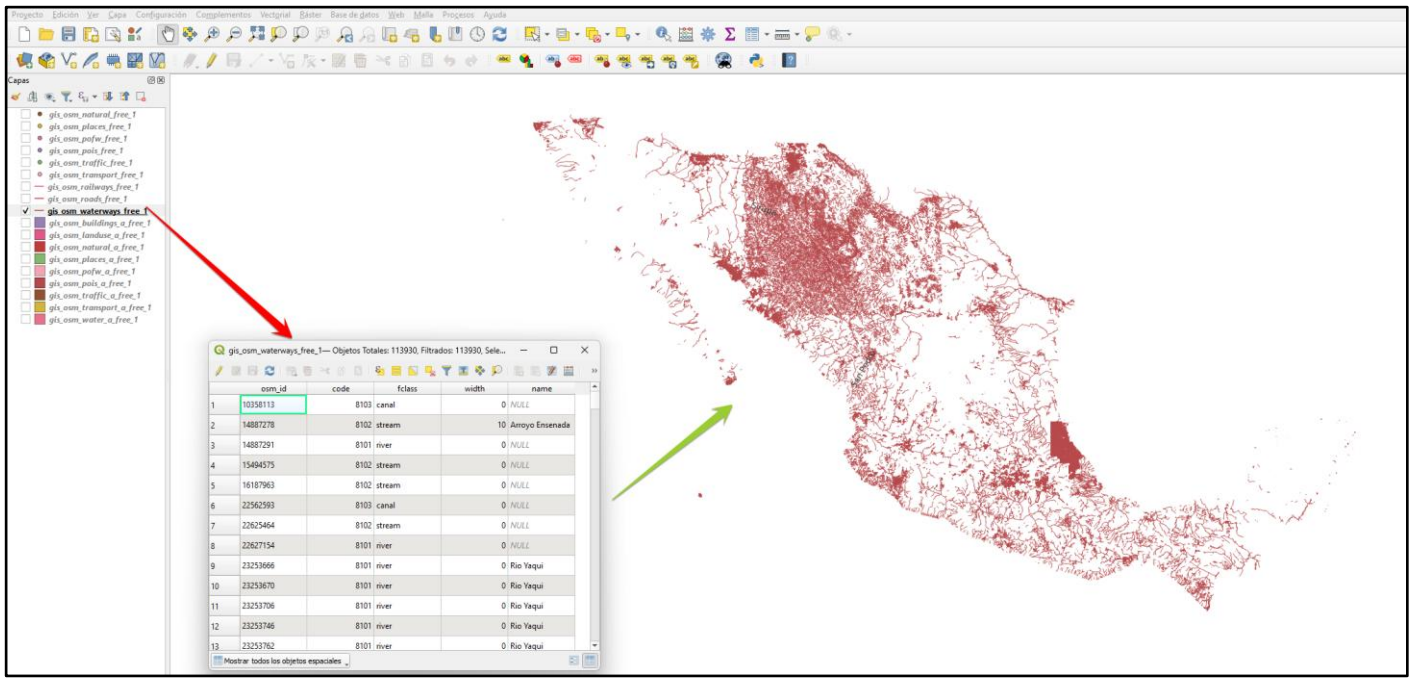
Paso 1: Explorando Shapefiles con QGIS:

Para examinar los datos de los Shapefiles usaré la **aplicación QGIS** para visualizar los datos geográficos, así como sus tablas y registros.

Se observan los siguientes archivos:

1. **gis_osm_natural_free_1:** Datos relacionados con características naturales (por ejemplo, ríos, lagos, montañas).
2. **gis_osm_places_free_1:** Información sobre lugares (por ejemplo, ciudades, pueblos).
3. **gis_osm_pofw_free_1:** Puntos de culto o lugares de adoración (por ejemplo, iglesias, mezquitas).
4. **gis_osm_pois_free_1:** Puntos de interés (por ejemplo, restaurantes, hoteles).
5. **gis_osm_traffic_free_1:** Datos relacionados con el tráfico (por ejemplo, semáforos, señales).
6. **gis_osm_transport_free_1:** Información sobre transporte (por ejemplo, paradas de autobús, estaciones de tren).
7. **gis_osm_railways_free_1:** Datos sobre vías férreas.
8. **gis_osm_roads_free_1:** Información sobre carreteras y calles.
9. **gis_osm_waterways_free_1:** Datos sobre vías fluviales (por ejemplo, ríos, canales).
10. **gis_osm_buildings_a_free_1:** Edificios.
11. **gis_osm_landuse_a_free_1:** Uso del suelo (por ejemplo, áreas residenciales, industriales).
12. **gis_osm_natural_a_free_1:** Características naturales con más detalle.
13. **gis_osm_places_a_free_1:** Lugares con más detalle.
14. **gis_osm_pofw_a_free_1:** Puntos de culto con más detalle.
15. **gis_osm_pois_a_free_1:** Puntos de interés con más detalle.
16. **gis_osm_traffic_a_free_1:** Datos de tráfico con más detalle.
17. **gis_osm_transport_a_free_1:** Información de transporte con más detalle.
18. **gis_osm_water_a_free_1:** Datos relacionados con agua (por ejemplo, lagos, océanos).

Por ejemplo, en el apartado de capas, seleccioné el archivo **gis_osm_waterways_free_1**, y se visualizan los datos del archivo y también el gráfico respectivo, así, explorando todo, me doy una idea de todo lo que contiene todos los Shapefiles.



Paso 2: Explorando datos de GPS:

Los datos de las rutas de los taxis en el CSV contienen las siguientes columnas:

- **id:** Identificador único del viaje.
- **vendor_id:** Empresa de taxi.
- **pickup_datetime:** Fecha y hora de Inicio.
- **dropoff_datetime:** Fecha y hora de Fin.
- **pickup_longitude:** Longitud de Inicio.
- **pickup_latitude:** Latitud de Inicio.
- **dropoff_longitude:** Longitud de Fin.
- **dropoff_latitude:** Latitud de Fin.
- **store_and_fwd_flag:** Indicador de almacenamiento y reenvío de datos.
- **trip_duration:** Duración del viaje en segundos.
- **dist_meters:** Distancia recorrida en metros.
- **wait_sec:** Tiempo de espera en segundos.

Paso 3: Preparar la Base de datos y las Tablas:

Ahora creé una base de datos llamada "**DATOS_GEOGRAFICOS**", donde colocaré todos los archivos del **Shapefiles**, cada uno en una tabla diferente y también una tabla con los **datos del GPS**.

a) Para los datos de los Shapefiles creé 18 tablas según lo siguiente:

Shapefile	Table in DataBase
gis_osm_natural_free_1	Natural
gis_osm_buildings_a_free_1	Edificio
gis_osm_landuse_a_free_1	Suelo
gis_osm_natural_a_free_1	Natural_mas_Detalle
gis_osm_traffic_free_1	Trafico

gis_osm_traffic_a_free_1	Trafico_mas_Detalle
gis_osm_railways_free_1	Vias_Ferreas
gis_osm_roads_free_1	Carretera_Calle
gis_osm_waterways_free_1	Vias_Fluviales
gis_osm_places_free_1	Lugares
gis_osm_places_a_free_1	Lugares_mas_Detalle
gis_osm_pois_free_1	Interes
gis_osm_pois_a_free_1	Interes_mas_Detalle
gis_osm_pofw_free_1	Culto
gis_osm_pofw_a_free_1	Culto_mas_Detalle
gis_osm_transport_free_1	Transporte
gis_osm_transport_a_free_1	Transporte_mas_Detalle
gis_osm_water_a_free_1	Agua

```sql
/*-----------1. Tabla Natural------------*/
CREATE TABLE Natural (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);


/*-----------2. Tabla Edificio------------*/
CREATE TABLE Edificio (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Type NVARCHAR(100),
    Geometria GEOMETRY
);


/*-----------3. Tabla Suelo------------*/
CREATE TABLE Suelo (
    Osm_id BIGINT,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*----4. Tabla Natural mas Detalles----*/
CREATE TABLE Natural_mas_Detalle (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------5. Trafico------------*/
CREATE TABLE Trafico (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);
```

```sql
/*-----------6. Trafico mas Detalle------------*/
CREATE TABLE Trafico_mas_Detalle (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------7. Vias_Ferreas------------*/
CREATE TABLE Vias_Ferreas (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Layer INT,
    Bridge NVARCHAR(5),
    Tunnel NVARCHAR(5),
    Geometria GEOMETRY
);

/*-----------8. Carreteras_Calles------------*/
CREATE TABLE Carreteras_Calles (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Ref NVARCHAR(100),
    Oneway NVARCHAR(5),
    MaxSpeed INT,
    Layer INT,
    Bridge NVARCHAR(5),
    Tunnel NVARCHAR(5),
    Geometria GEOMETRY
);


/*-----------9. Vias_Fluviales------------*/
CREATE TABLE Vias_Fluviales (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Width INT,
    Name NVARCHAR(100),
    Geometria GEOMETRY
);
```

```sql
/*-----------10. Lugares------------*/
CREATE TABLE Lugares (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Population INT,
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------11. Lugares_mas_Detalle------------*/
CREATE TABLE Lugares_mas_Detalle (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Population INT,
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------12. Interes------------*/
CREATE TABLE Interes (
    Osm_id BIGINT NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------13. Interes_mas_Detalle------------*/
CREATE TABLE Interes_mas_Detalle (
    Osm_id BIGINT NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);

/*-----------14. Culto------------*/
CREATE TABLE Culto (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);
```

```sql
/*-----------15. Culto_mas_Detalle------------*/
CREATE TABLE Culto_mas_Detalle (
    Osm_id BIGINT PRIMARY KEY NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);


/*-----------16. Transporte------------*/
CREATE TABLE Transporte (
    Osm_id BIGINT NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);
```

```sql
/*-----------17. Transporte_mas_Detalle------------*/
CREATE TABLE Transporte_mas_Detalle (
    Osm_id BIGINT NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);


/*-----------18. Agua------------*/
CREATE TABLE Agua (
    Osm_id BIGINT NOT NULL,
    Code INT,
    Fclass NVARCHAR(100),
    Name NVARCHAR(100),
    Geometria GEOMETRY
);
```

b) Para los datos de GPS lo inserté en una tabla agregándole dos columnas más, llamadas **UbicacionWKT_Recogida** y **UbicacionWKT_Entrega**, estas columnas representan

```sql
CREATE TABLE TaxiViajes_GPS (
    ID INT PRIMARY KEY,
    Proveedor NVARCHAR(50),
    FechaHora_Recogida DATETIME,
    FechaHora_Entrega DATETIME,
    Longitud_Recogida FLOAT,
    Latitud_Recogida FLOAT,
    Longitud_Entrega FLOAT,
    Latitud_Entrega FLOAT,
    Indicador_Almacenamiento_Reenvio NVARCHAR(10),
    DuracionViaje_Segundos INT,
    Distancia_Metros INT,
    TiempoEspera_Segundos BIGINT,
    UbicacionWKT_Recogida NVARCHAR(200),
    UbicacionWKT_Entrega NVARCHAR(200)
);
```

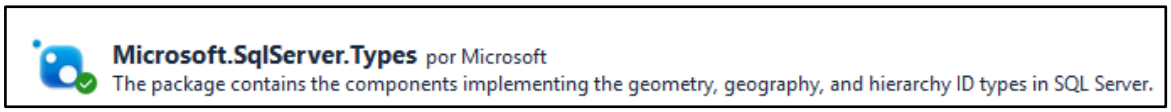
Paso 4: Insertando Datos de Shapefiles a la Base de Datos con SSIS.

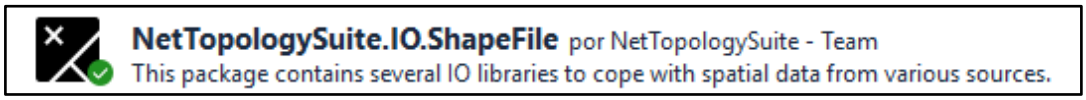
Para este paso, crearé un proyecto en SSIS y como no existe un componente nativo para insertar Shapefiles(.shp) directamente a SQL Server, emplearé **código en C# (Aplicación de Consola (.NET Framework) C#)**, donde usaré librerías para la correcta inserción de datos hacia la BD en SQL Server.

Por ejemplo, para el Shapefile Agua, el código en C# sería:

```csharp
using System;
using System.Data;
using System.Data.SqlClient;
using System.Text;
using Microsoft.SqlServer.Types;
using NetTopologySuite.Geometries;
using NetTopologySuite.IO;

namespace _18.Agua
{
    0 referencias
    class Program
    {
        0 referencias
        static void Main(string[] args)
        {
            string shapefilePath = @"D:\2. PORTAFOLIO-MAXIMO-SILVA\SQL Server Integration Services (SSIS)\ANALISIS DE DATOS GEOGRAFICOS\Dataset\Shapefiles\gis_osm_water_a_free_1.shp"; // Ruta del Shapefile

            string connectionString = "Server=MAX\\MSSQLSERVER2022;Database=DATOS_GEOGRAFICOS;User Id=sa;Password=123456789;";

            try
            {
                using (var reader = new ShapefileDataReader(shapefilePath, GeometryFactory.Default, Encoding.UTF8))
                using (SqlConnection conn = new SqlConnection(connectionString))
                {
                    conn.Open();
                    int totalRegistros = 0;

                    while (reader.Read())
                    {
                        long osm_id = Convert.ToInt64(reader["osm_id"]);
                        int code = Convert.ToInt32(reader["code"]);
                        string fclass = reader["fclass"]?.ToString() ?? "";
                        string name = reader["name"]?.ToString() ?? "";

                        Geometry geometry = reader.Geometry;
                        SqlGeometry sqlGeom = SqlGeometry.Null;

                        if (geometry != null)
                        {
                            // Convertir NetTopologySuite Geometry a WKB para SQL Server
                            WKBWriter wkbWriter = new WKBWriter();
                            byte[] wkbBytes = wkbWriter.Write(geometry);
                            sqlGeom = SqlGeometry.STGeomFromWKB(new System.Data.SqlTypes.SqlBytes(wkbBytes), 4326); // 4326 = SRID para coordenadas geográficas
                        }

                        // Insertar en SQL Server
                        using (SqlCommand cmd = new SqlCommand("INSERT INTO Agua (Osm_id, Code, Fclass, Name, Geometria) VALUES (@osm_id, @code, @fclass, @name, @geom)", conn))
                        {
                            cmd.Parameters.Add("@osm_id", SqlDbType.BigInt).Value = osm_id;
                            cmd.Parameters.Add("@code", SqlDbType.Int).Value = code;
                            cmd.Parameters.Add("@fclass", SqlDbType.NVarChar, 100).Value = fclass;
                            cmd.Parameters.Add("@name", SqlDbType.NVarChar, 100).Value = name;


                            // Aquí se establece el UDTTypeName para SqlGeometry
                            SqlParameter geomParam = cmd.Parameters.Add("@geom", SqlDbType.Udt);
                            geomParam.Value = sqlGeom;
                            geomParam.UdtTypeName = "Geometry"; // Nombre del tipo UDT en SQL Server

                            cmd.ExecuteNonQuery();
                        }
                        Console.WriteLine($"Insertado: osm_id={osm_id}, geom={sqlGeom.ToString()}");
                        totalRegistros++;
                    }

                    Console.WriteLine("-------------------------------------------------");
                    Console.WriteLine($"Total de registros insertados: {totalRegistros}");
                }
            }
            catch (Exception ex)
            {
                Console.WriteLine($"Error: {ex.Message}");
            }
        }
    }
}
```

Para que funcione el código, se necesitan instalar dos paquetes dentro **del Administrador de Paquetes de NuGet:**

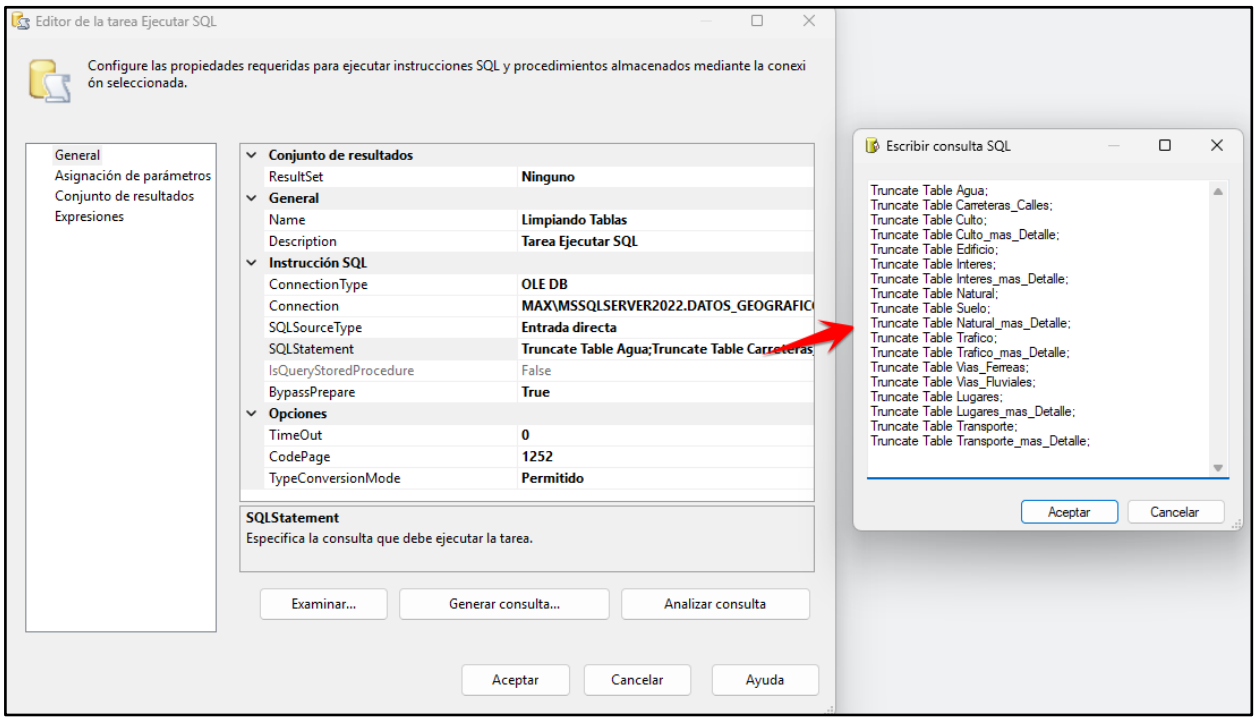




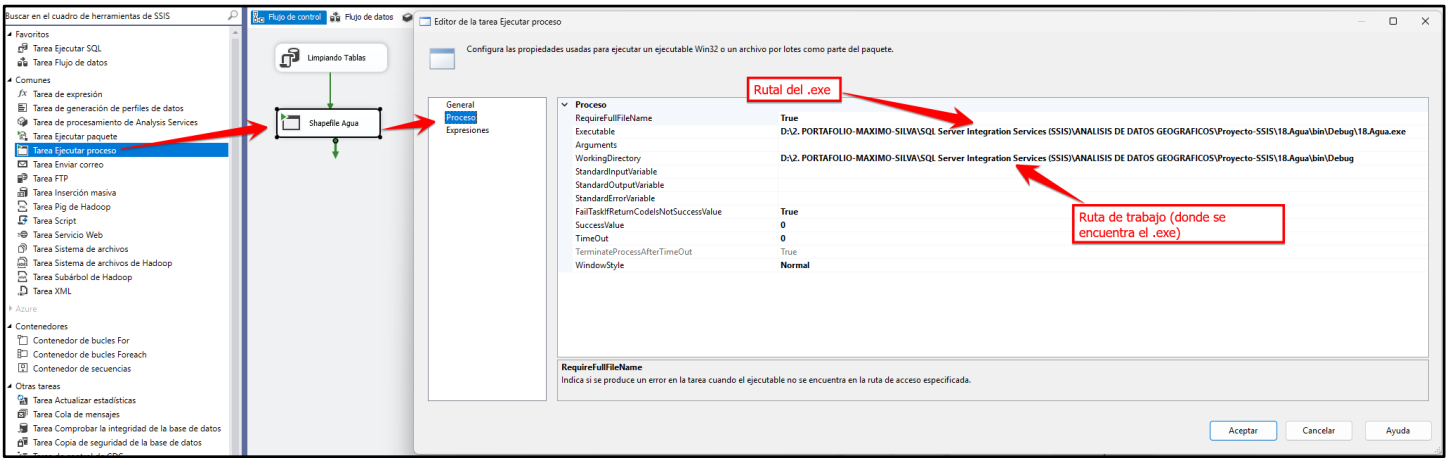
Una vez todo funcione correctamente, compilé el proyecto para usarlo posteriormente.

El mismo proceso se realizó para todos los archivos shapefiles.

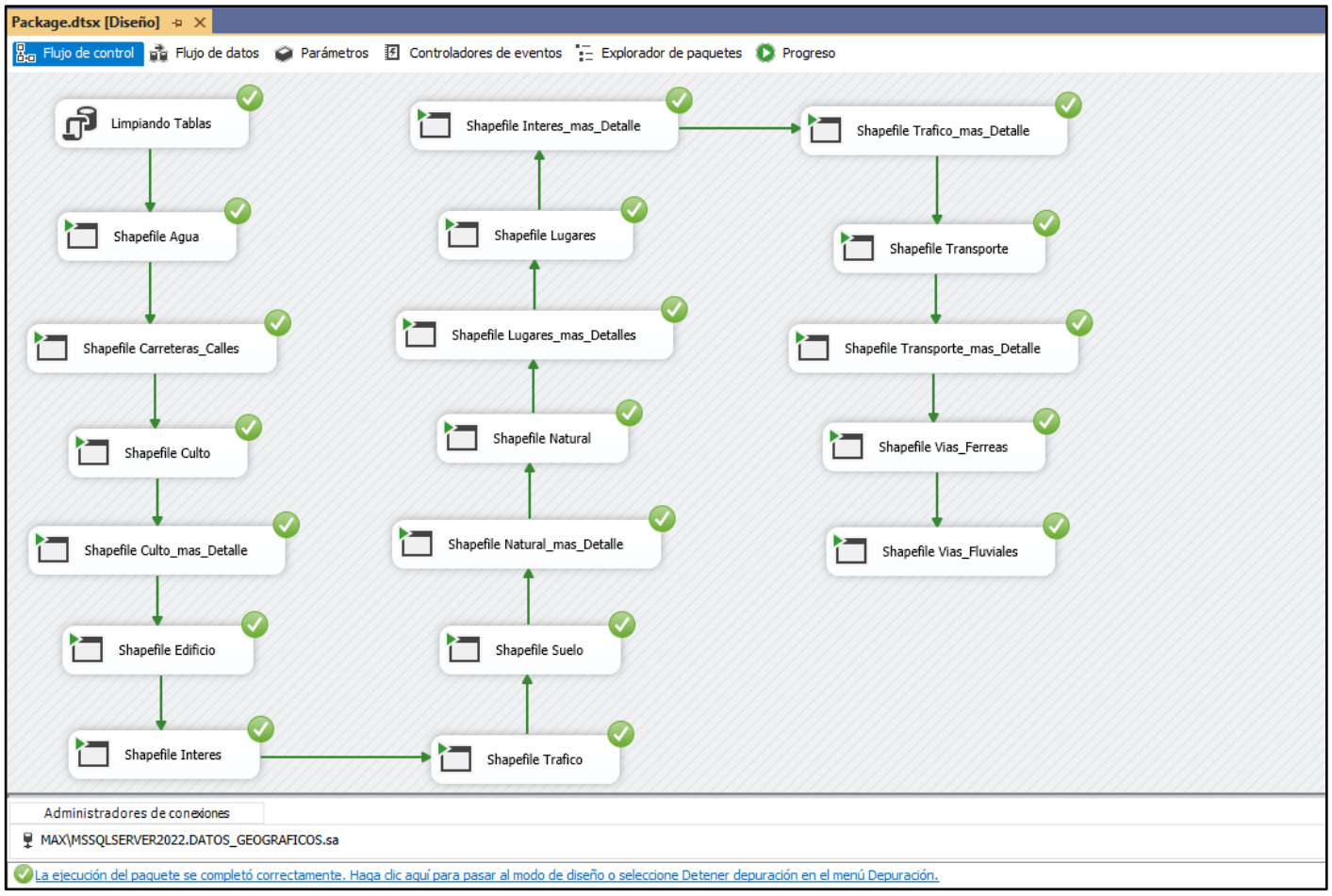
Después arrastré el componente de Tarea SQL para que realice en primer lugar el truncamiento de todas las tablas.



Luego empleé la tarea de Ejecutar Paquete donde configuré la ruta del ejecutable que generó la compilación:

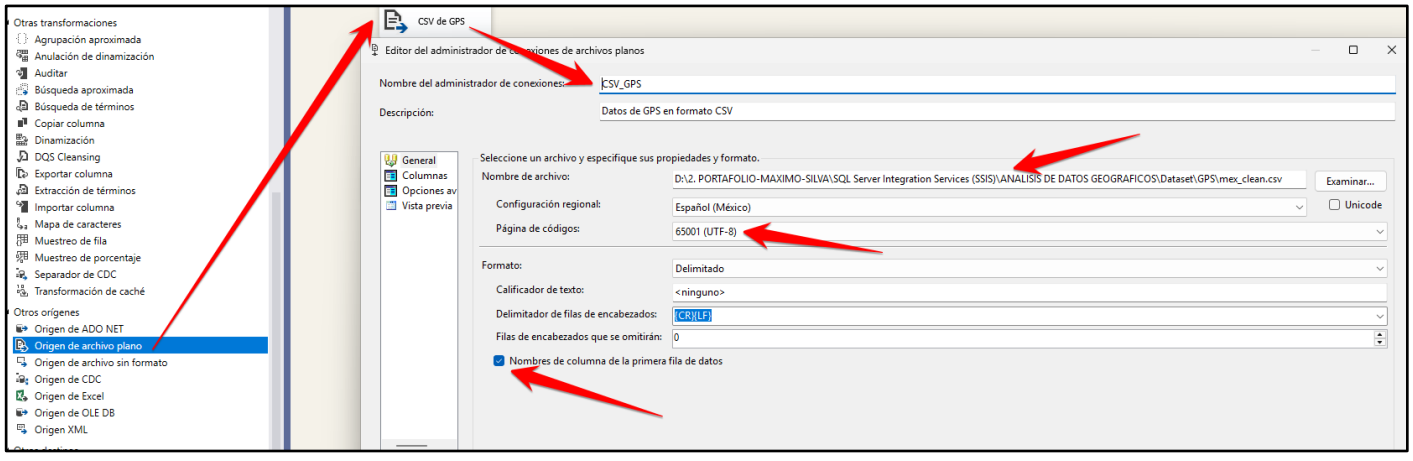


Lo mismo realicé para todos los archivos de cada Shapefile y ejecuté el proyecto:

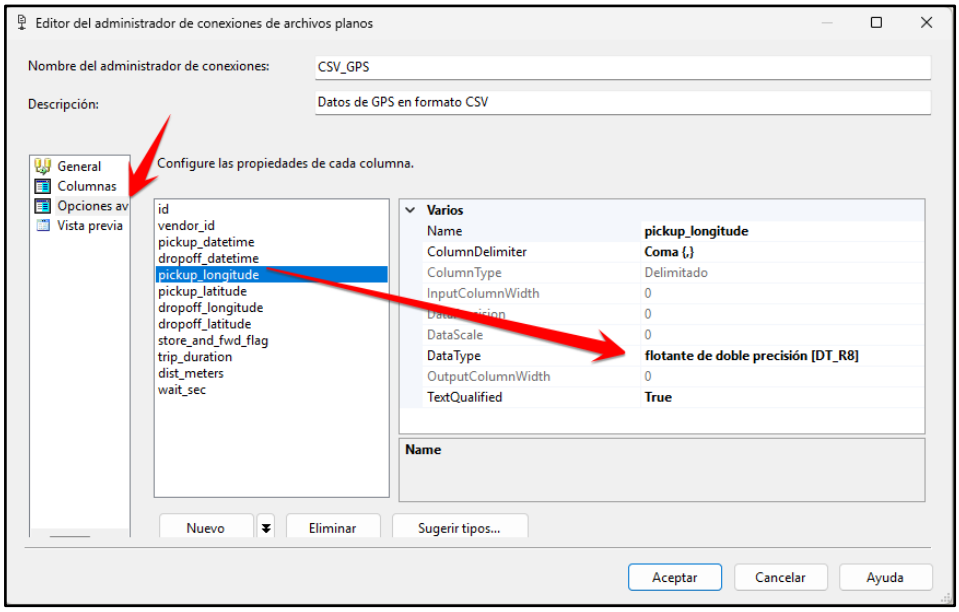


Paso 5: Insertando Datos del GPS a la Base de Datos con SSIS.

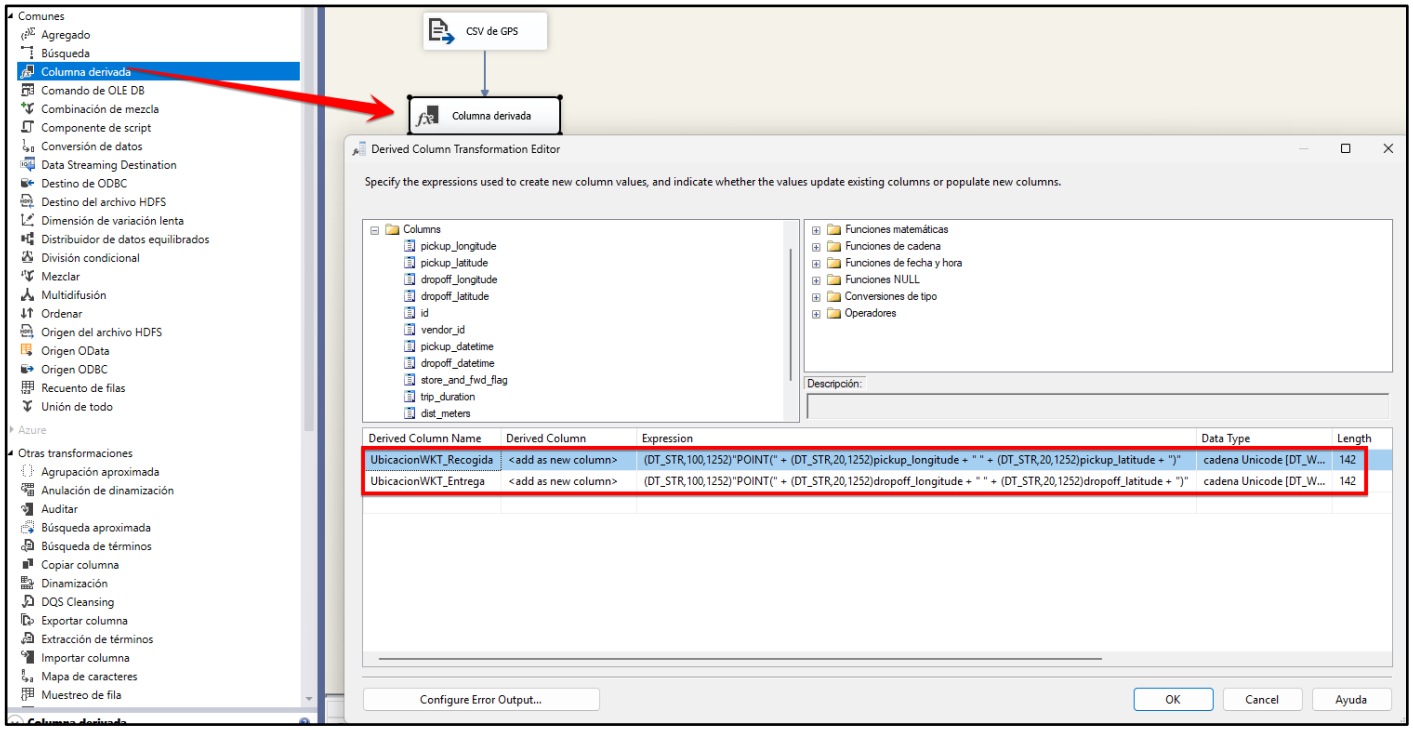
Arrastré un origen de archivo plano y lo configuré con el .CSV que contiene los datos del GPS.



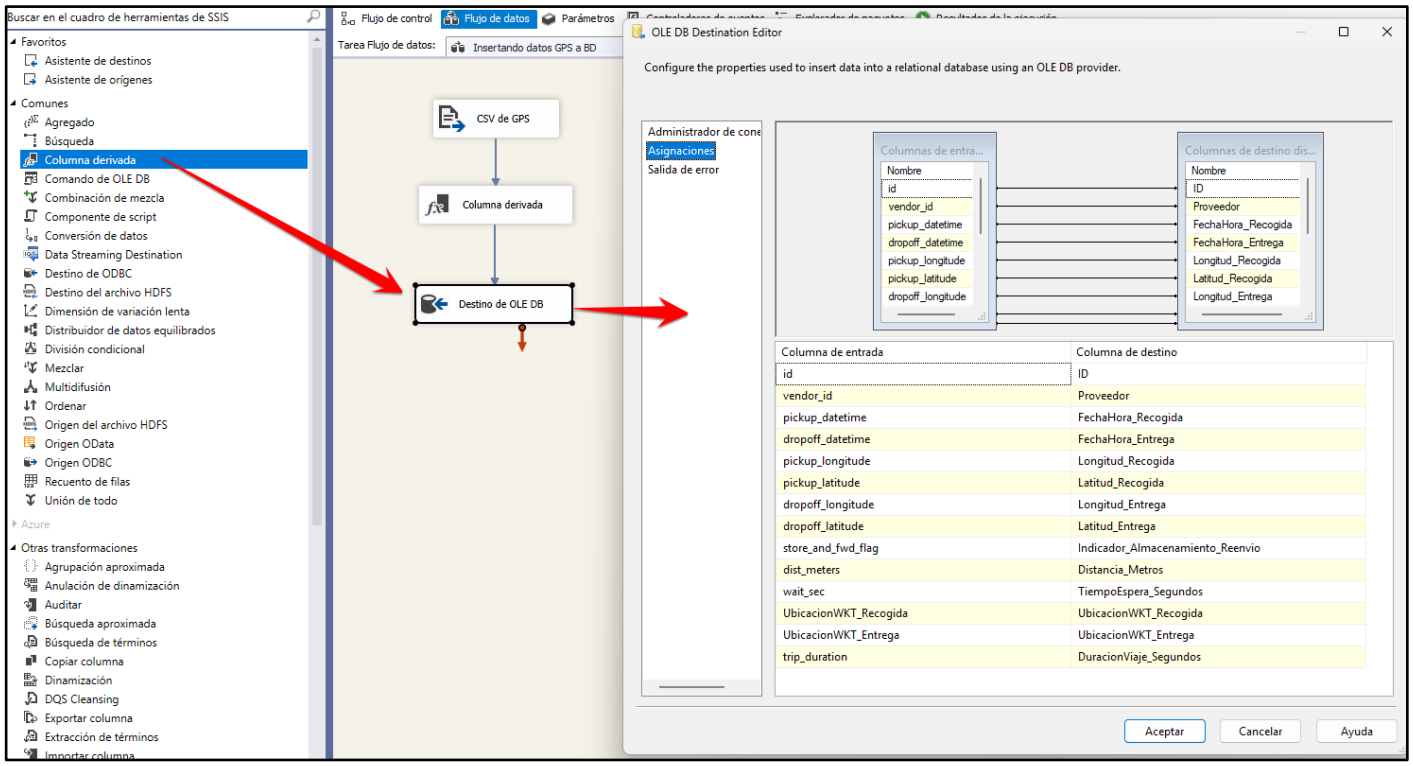
Se coloca en cada columna el tipo correspondiente en SSIS teniendo en cuenta su tipo de dato en SQL Server. Por ejemplo, **pickup_longitude** es de tipo **DT_R8** en SSIS para que pueda guardarse de forma correcta en SQL Server donde el tipo es FLOAT.

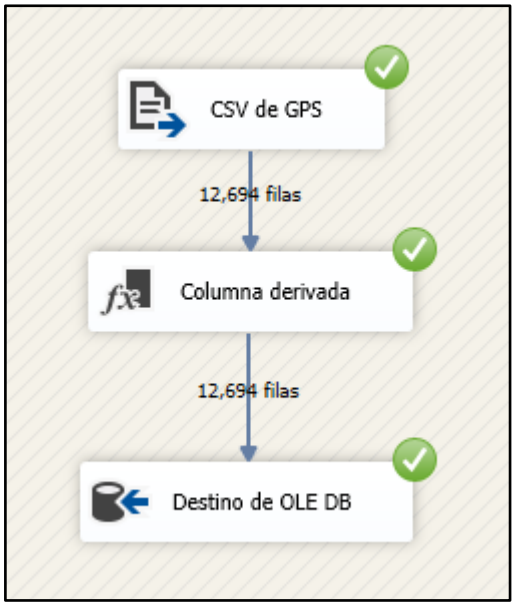


Luego empleé el componente de Columna Derivada y se creó dos columnas que guardarán en una cadena de texto el formato de WKT para más adelante poder convertirlo a un tipo de datos GEOGRAPHY.



Por último, configuré un destino hacia una tabla en SQL Server y ejecuto el proyecto.





Paso 6: Identificando Datos a utilizarse para análisis.

Como los datos ingresados en la primera base de datos no los utilizaré todos, además algunos datos son incoherentes, haré un proceso de ETL hacia otra base de datos con solo las columnas que utilizaré y utilizando datos limpios. Para ello creé una base de datos llamada "DATOS_GEOGRAFICO_LIMPIOS", ahí almacenaré todas las tablas que se utilizarán para diversos análisis.

Se desea exportar datos que resuelvan las siguientes interrogantes:

a) ¿Cuántos taxis recogen pasajeros cerca de puntos de interés (hoteles, restaurantes, estaciones de transporte, parques, ¿etc.)?

b) ¿Qué porcentaje de los viajes comienzan o terminan cerca de una estación de transporte?

c) ¿Cuáles son las zonas con más viajes de taxi en una hora específica del día?

d) ¿Qué punto de interés es el que contiene mayor cantidad de viajes registrados entre Hoteles y Restaurantes?

e) ¿Cuáles son las carreteras más utilizadas para iniciar un viaje?

f) ¿Cuáles son los lugares de interés con más viajes de taxi?

g) ¿Cuáles son las ciudades con mayor actividad de taxis?

h) Listado de viajes que empiezan cerca de algún área natural.

i) ¿Cuál es la velocidad promedio de los viajes en diferentes tipos de vía en Km/h?

j) ¿Cuántos viajes terminan en Hospitales?

Las tablas de origen de los Shapefiles con datos útiles que servirán para este análisis son:

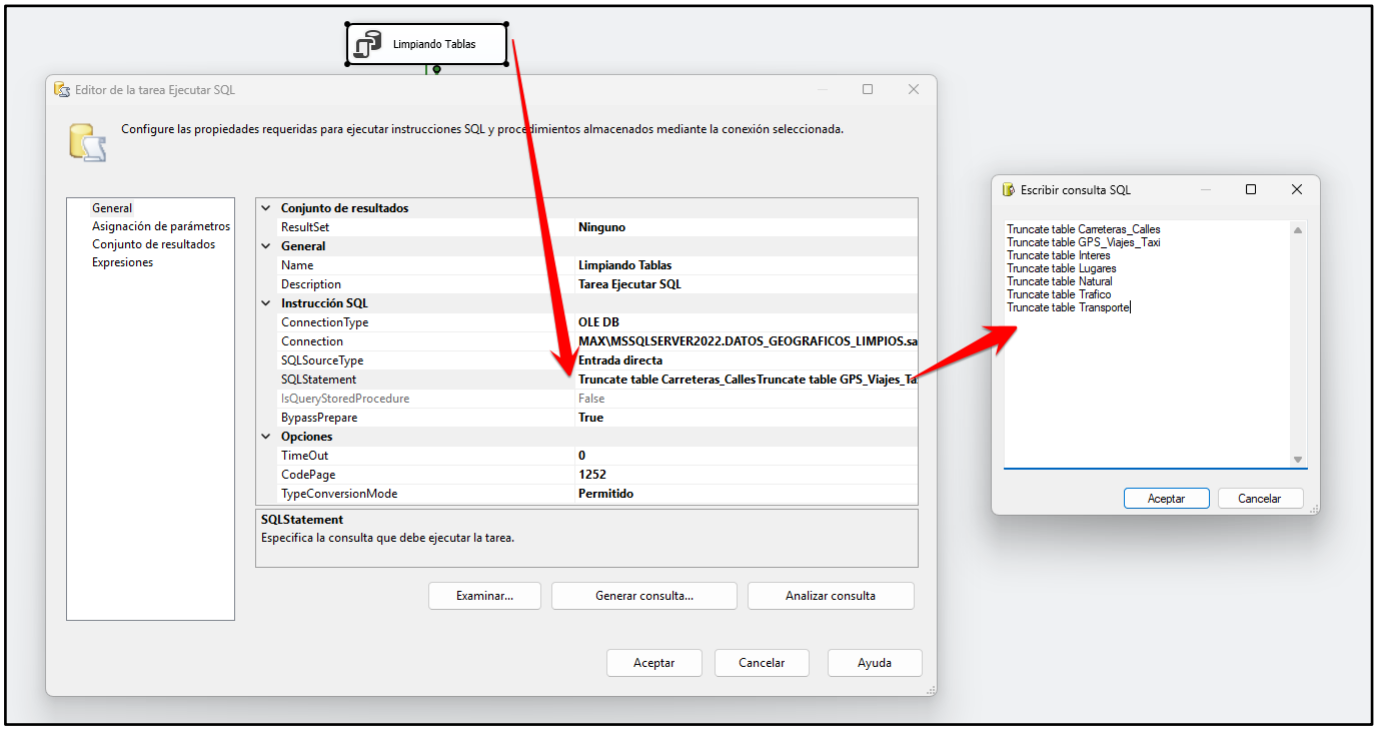
TABLAS EN BD DATOS_GEOGRAFICOS	DESCRIPCIÓN Y UTILIDAD	TABLAS EN BD DATOS_GEOGRAFIOS_LIMPIOS
Carretera_Calle	Clave para analizar la infraestructura vial y los trayectos	**Carretera_Calle**
Interes	Datos de hoteles, restaurantes, estaciones de tren, etc.	**Interes**
Lugares	Puede ayudar a ver puntos clave de Inicio y fin del viaje.	**Lugares**
Lugares_mas_Detalle	Más datos de Lugares clave.	**Lugares**
Agua	Datos de lagos, ríos, etc.	**Natural**
Natural	Datos referentes a áreas naturales.	**Natural**
Natural_mas_Detalle	Más datos de área Naturales.	**Natural**
Trafico	Útil para analizar congestión y tiempo de viaje	**Trafico**
Trafico_mas_Detalle	Más datos referentes a tráficos.	**Trafico**
Transporte	Paradas de autobús, estación de tren, aeropuerto.	**Transporte**
Transporte_mas_Detalle	Más datos de Transportes.	**Transporte**

Y las columnas de los datos del GPS que se usarán son:

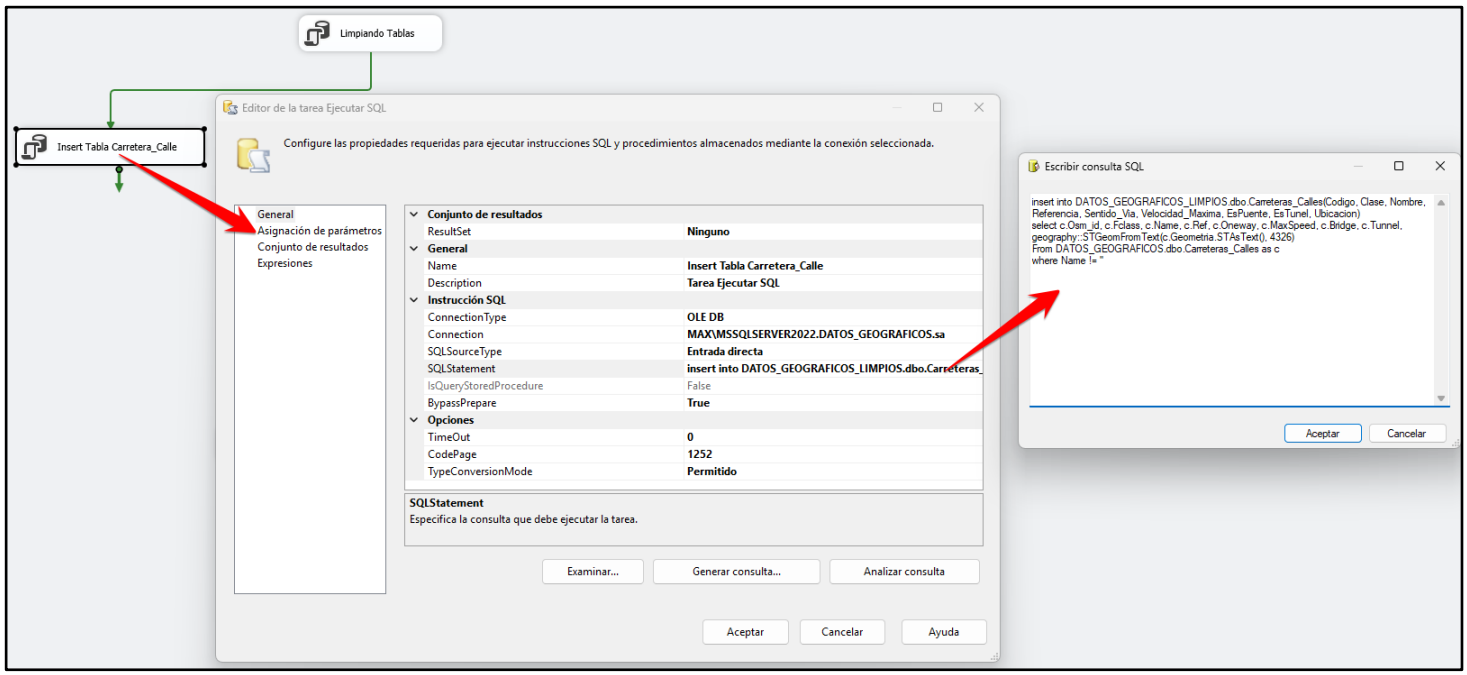
COLUMNAS DE LA TABLA TaxiViajes_GPS BD DATOS_GEOGRAFICOS	DESCRIPCIÓN	COLUMNAS DE LA TABLA GPS_Viajes_Taxi BD DATOS_GEOGRAFIOS_LIMPIOS
ID	Código	**Id_Viajes_Taxi**
Proveedor	Proveedor de Taxis	**Proveedor**
FechaHora_Entrega	Fecha y Hora de Inicio del viaje	**FechaHora_Inicio**
FechaHora_Recogida	Fecha y Hora de Fin del viaje	**FechaHora_Fin**
DuracionViaje_Segundos	Duración del viaje en segundos	**DuracionViaje_Segundos**
Distancia_Metros	Distancia en metros del viaje	**Distancia_Metros**
TiempoEspera_Segundos	Tiempo de espera en segundos	**TiempoEspera_Segundos**
UbicacionWKT_Recogida (NVARCHAR)	Ubicación de Inicio del Viaje	**Ubicacion_Inicio (GEOGRAPHY)**
UbicacionWKT_Entrega (NVARCHAR)	Ubicación del destino del Viaje	**Ubicacion_Fin (GEOGRAPHY)**

Paso 7: Pasando datos de los Shapefiles a una nueva BD solo con datos filtrados.

En primer lugar, hice un truncamiento de las tablas utilizando el componente Tarea Ejecutar SQL:



Debido a que el tipo de dato Geometría se maneja de una manera compleja en SSIS, opte por realizar cada inserción de datos usando también el componente de Tarea Ejecutar SQL, **convirtiendo los datos GEOMETRY a GEOGRAPHY** y haciendo filtros, como pasando solo datos que contengan datos en el campo de Nombre, tal cual la siguiente imagen:

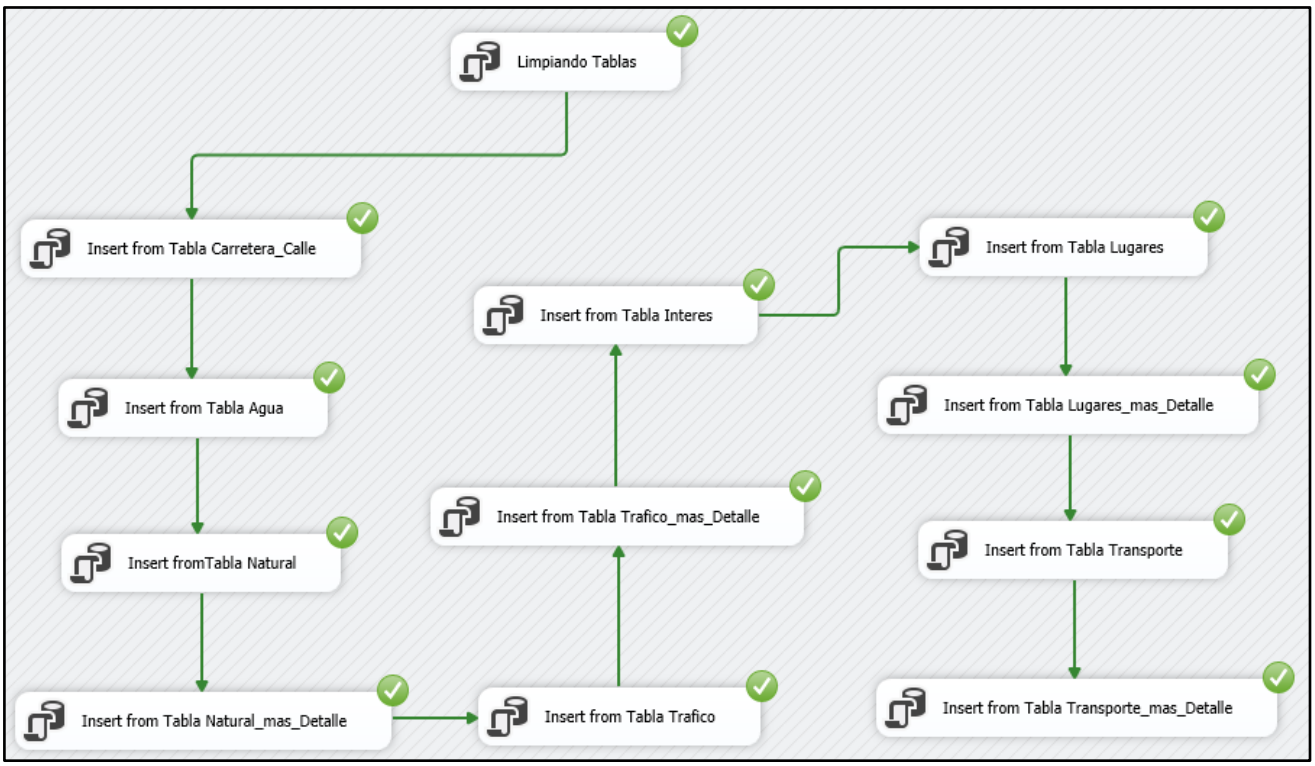


El mismo proceso se realizó para todas las tablas, utilizando las siguientes consultas:

```
/*-------------INSERT CARRETERA_CALLE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Carreteras_Calles(Codigo, Clase, Nombre, Referencia, Sentido_Via,
Velocidad_Maxima, EsPuente, EsTunel, Ubicacion)
select c.Osm_id, c.Fclass, c.Name, c.Ref, c.Oneway, c.MaxSpeed, c.Bridge, c.Tunnel,
geography::STGeomFromText(c.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Carreteras_Calles as c
where Name != ''



/*-------------INSERT AGUA----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Natural(Codigo, Clase, Nombre, Ubicacion)
select a.Osm_id, a.Fclass, a.Name, geography::STGeomFromText(a.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Agua as a
where a.Name != ''


/*-------------INSERT NATURAL----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Natural(Codigo, Clase, Nombre, Ubicacion)
select l.Osm_id, l.Fclass, l.Name, geography::STGeomFromText(l.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Natural as l
where l.Name != ''

/*-------------INSERT NATURAL_MAS_DETALLE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Natural(Codigo, Clase, Nombre, Ubicacion)
select n.Osm_id, n.Fclass, n.Name, geography::STGeomFromText(n.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Natural_mas_Detalle as n
where n.Name != ''


/*-------------INSERT INTERES----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Interes(Codigo, Clase, Nombre, Ubicacion)
select i.Osm_id, i.Fclass, i.Name, geography::STGeomFromText(i.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Interes as i
where i.Name != ''

/*-------------INSERT TRAFICO----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Trafico(Codigo, Clase, Nombre, Ubicacion)
select t.Osm_id, t.Fclass, t.Name, geography::STGeomFromText(t.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Trafico as t


/*-------------INSERT TRAFICO_MAS_DETALLE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Trafico(Codigo, Clase, Nombre, Ubicacion)
select t.Osm_id, t.Fclass, t.Name, geography::STGeomFromText(t.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Trafico_mas_Detalle as t


/*-------------INSERT LUGARES----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Lugares(Codigo, Clase, Poblacion, Nombre, Ubicacion)
select l.Osm_id, l.Fclass, l.Population, l.Name, geography::STGeomFromText(l.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Lugares as l
where l.Name != ''
```

```
/*-------------INSERT LUGARES_MAS_DETALLE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Lugares(Codigo, Clase, Poblacion, Nombre, Ubicacion)
select l.Osm_id, l.Fclass, l.Population, l.Name, geography::STGeomFromText(l.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Lugares_mas_Detalle as l
where l.Name != ''


/*-------------INSERT TRANSPORTE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Transporte(Codigo, Clase, Nombre, Ubicacion)
select t.Osm_id, t.Fclass, t.Name, geography::STGeomFromText(t.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Transporte as t
where t.Name != ''


/*-------------INSERT TRANSPORTE_MAS_DETALLE----------------*/
insert into DATOS_GEOGRAFICOS_LIMPIOS.dbo.Transporte(Codigo, Clase, Nombre, Ubicacion)
select t.Osm_id, t.Fclass, t.Name, geography::STGeomFromText(t.Geometria.STAsText(), 4326)
From DATOS_GEOGRAFICOS.dbo.Transporte_mas_Detalle as t
where t.Name != ''
```

Se ejecuta el proyecto:

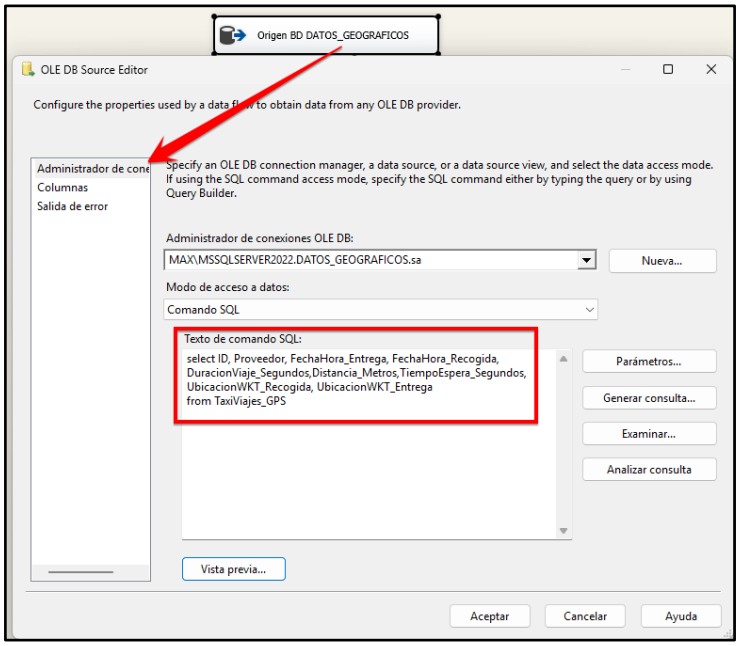


Paso 8: Pasando datos del GPS a una nueva BD solo con datos filtrados.

Para pasar los datos relevantes y haciendo los filtros correspondientes extraje los campos mencionados anteriormente, para ello realicé el siguiente proceso:

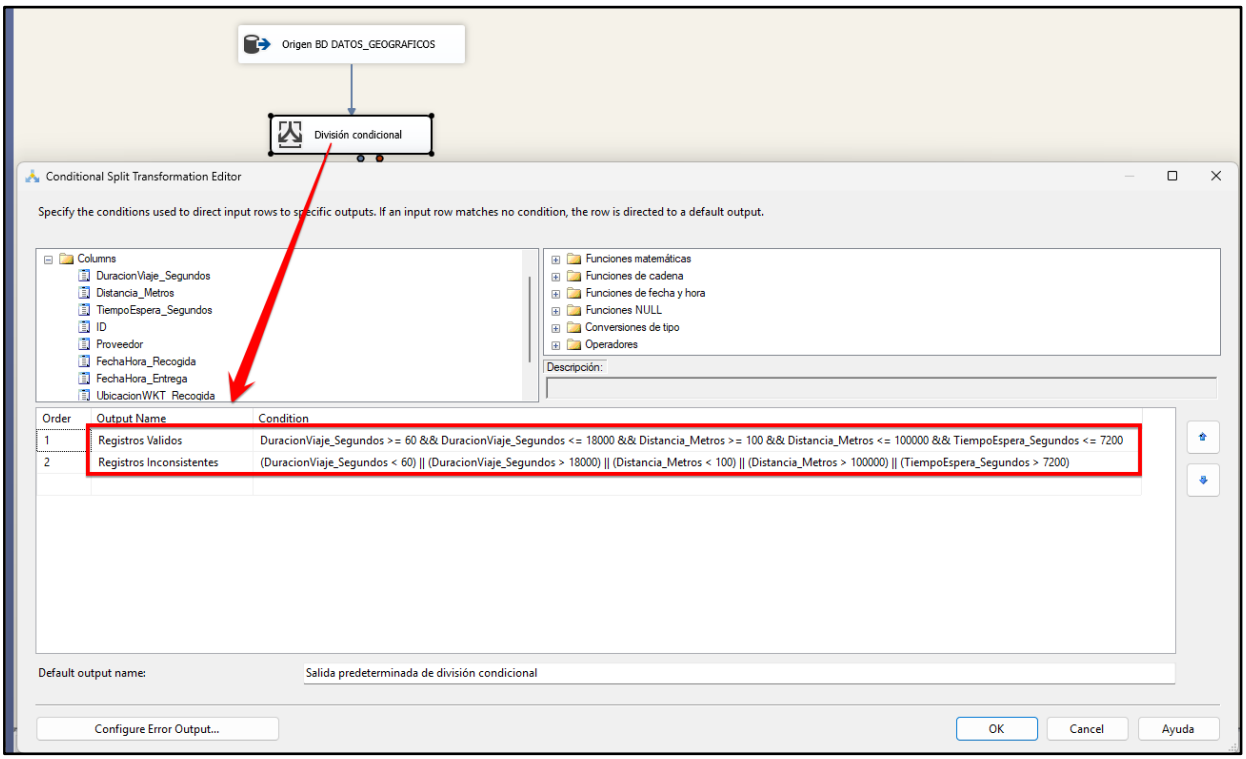
a) Hacer la consulta para extraer únicamente las columnas que requiero de la BD DATOS_GEOGRAFICOS, incluyendo los datos de la Ubicación de Inicio y Fin a campos NVARCHAR en la base de datos de destino, en las columnas de: Ubicacion_Inicio_Temporal y Ubicacion_Fin_Temporal.

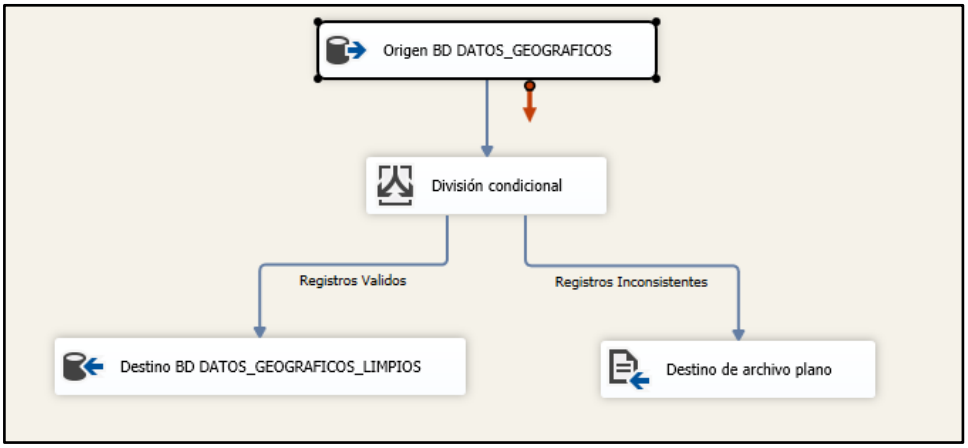
```
select ID, Proveedor, FechaHora_Entrega, FechaHora_Recogida,
DuracionViaje_Segundos,Distancia_Metros,TiempoEspera_Segundos,
geography::STGeomFromText(UbicacionWKT_Recogida, 4326) AS UbicacionRecogida,
geography::STGeomFromText(UbicacionWKT_Entrega, 4326) AS UbicacionEntrega
from DATOS_GEOGRAFICOS.dbo.TaxiViajes_GPS
```



b) Filtrar datos incoherentes utilizando el componente de División Condicional el cual usa los siguientes criterios:

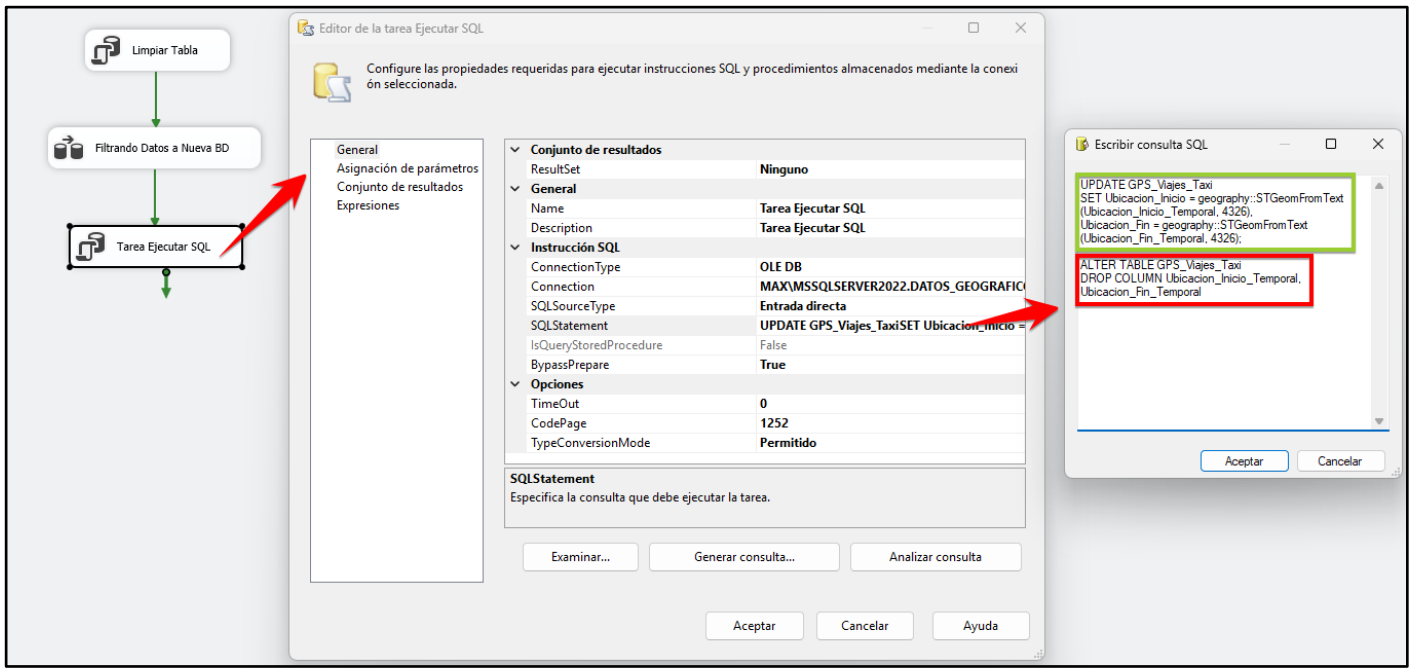
- **La duración del viaje debe ser mayor o igual a 60 segundos y menor de 5 horas (18000 segundos),** ya que no tiene lógica que los viajes duren pocos segundos, se trataría de un tipo de inconsistencia.
- **La distancia en metros debe ser mayor a 100 metros y menor a 100000 metros,** que probablemente fueron datos erróneos por la distancia indicada.
- **El tiempo de espera debe ser menor a 2 horas (7200 segundos)**, que, exagerando, es un tiempo el cual podría estar en espera para el inicio del viaje.
- **Los datos que están fuera de ese rango,** los pasaré a un .CSV para verificar que efectivamente tiene datos erróneos.



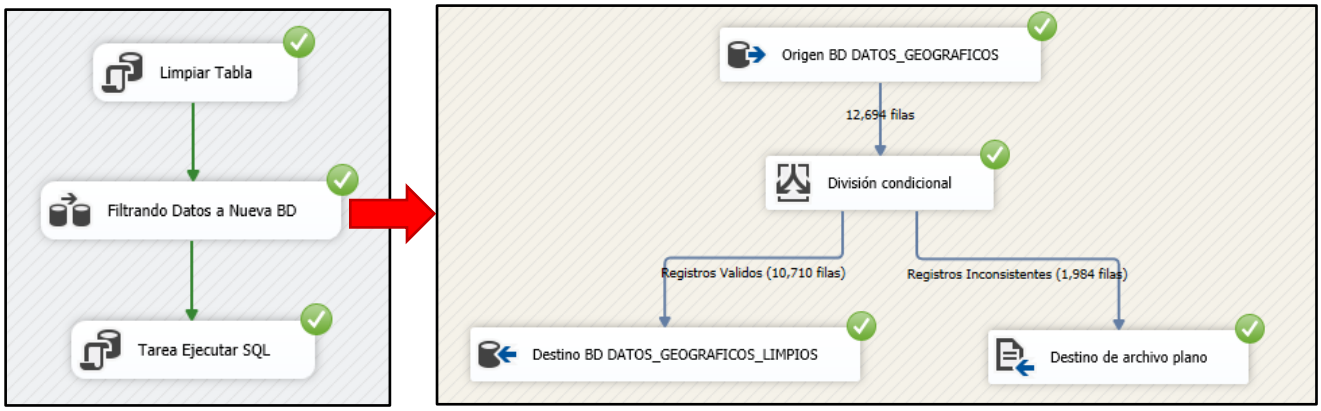


c) Convertir y pasar los datos de **Ubicacion_Inicio_Temporal** y **Ubicacion_Fin_Temporal** a **Ubicacion_Inicio** y **Ubicacion_Fin** respectivamente.

d) Eliminar las columnas Temporales: Ubicacion_Inicio_Temporal y Ubicacion_Fin_Temporal.



e) Ejecutar el proyecto.



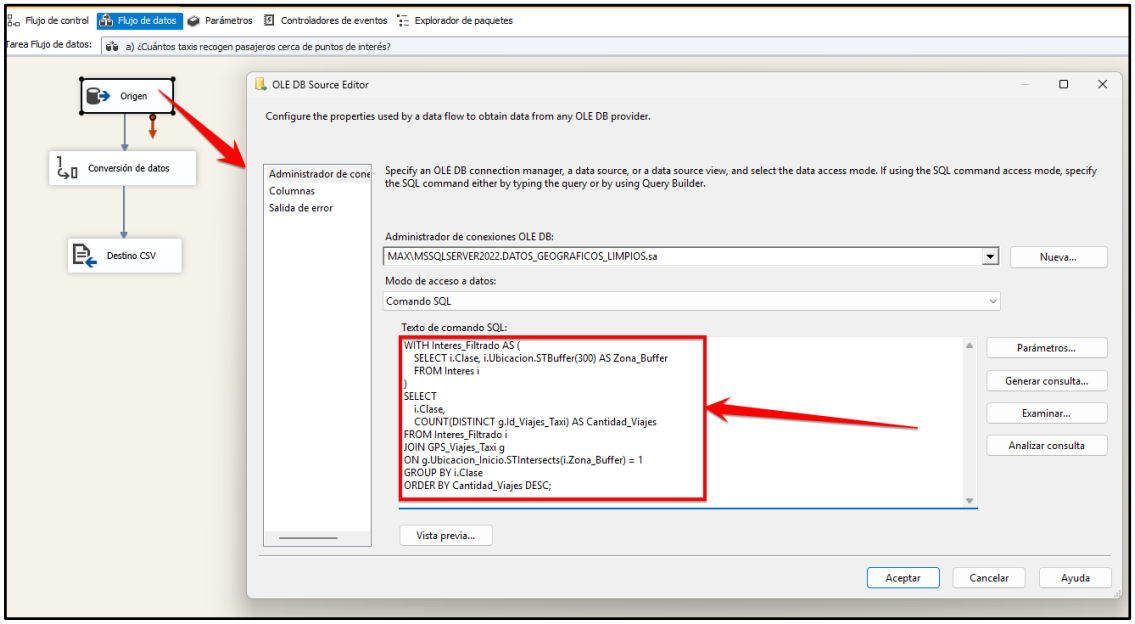
Paso 9: Realizando Análisis.

Para realizar el análisis de todas las preguntas se hizo una consultas en SQL Server, la cual extraerá los datos que respondan a cada análisis, luego, mediante un proceso de ETL en SSIS, esos datos los guardaré en un archivo de texto CSV.

a) ¿Cuántos taxis recogen pasajeros cerca de puntos de interés (hoteles, restaurantes, estaciones de transporte, parques, ¿etc.)?

Objetivo: Identificar cuántos taxis inician su viaje por cada grupo de interés.

Extrayendo datos en SSIS mediante consulta SQL:

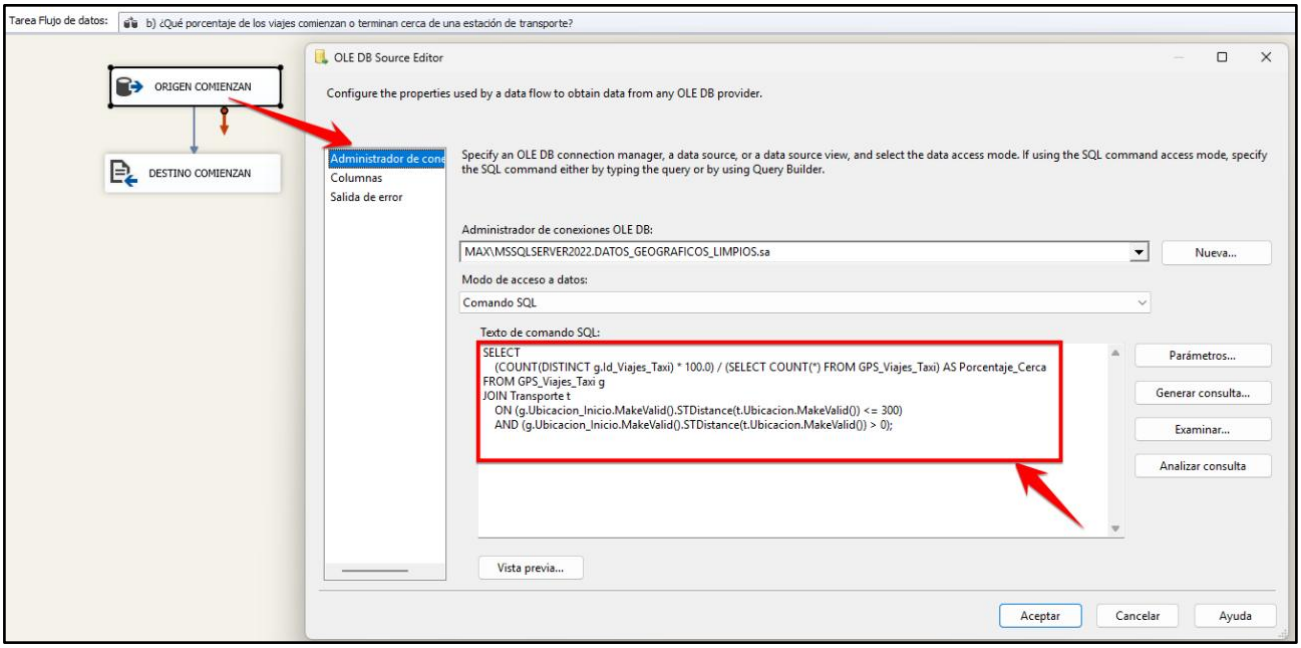


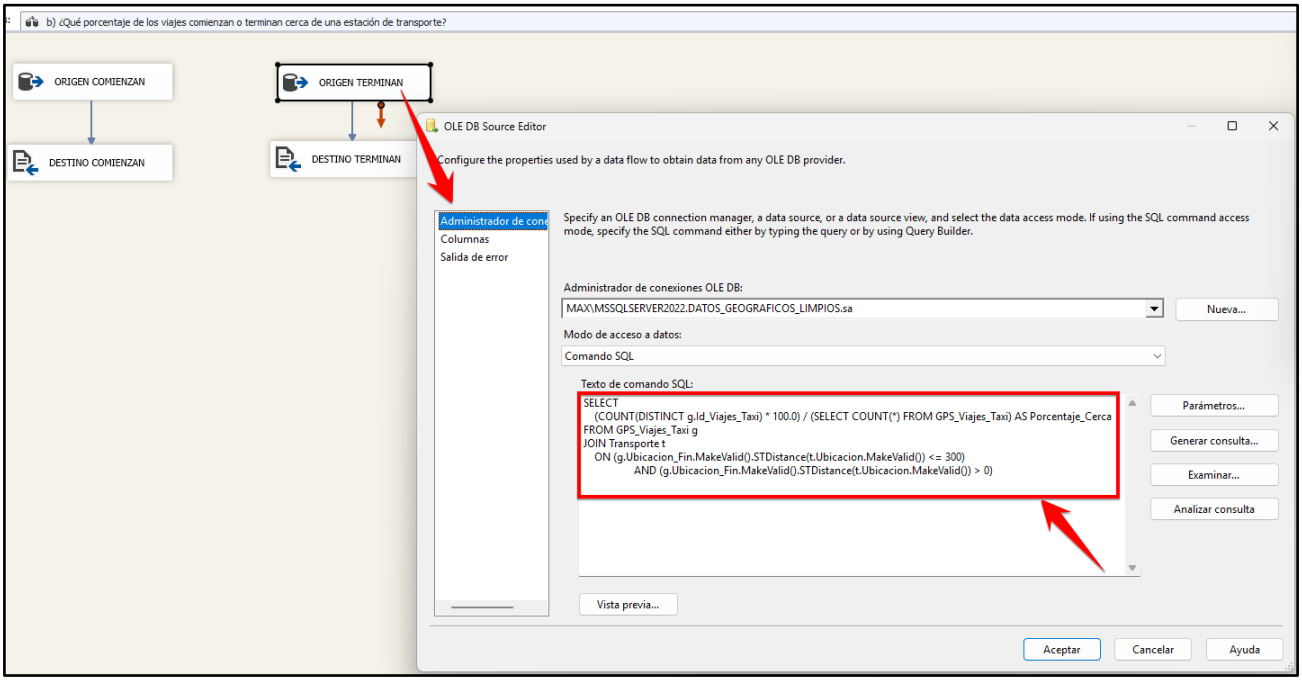
Destino: Un destino CSV.

b) ¿Qué porcentaje de los viajes comienzan o terminan cerca de una estación de transporte?

Objetivo: Identificar el porcentaje de viajes que inician o culminan cerca a cualquier estación de transporte.

Extrayendo datos en SSIS mediante consulta SQL:



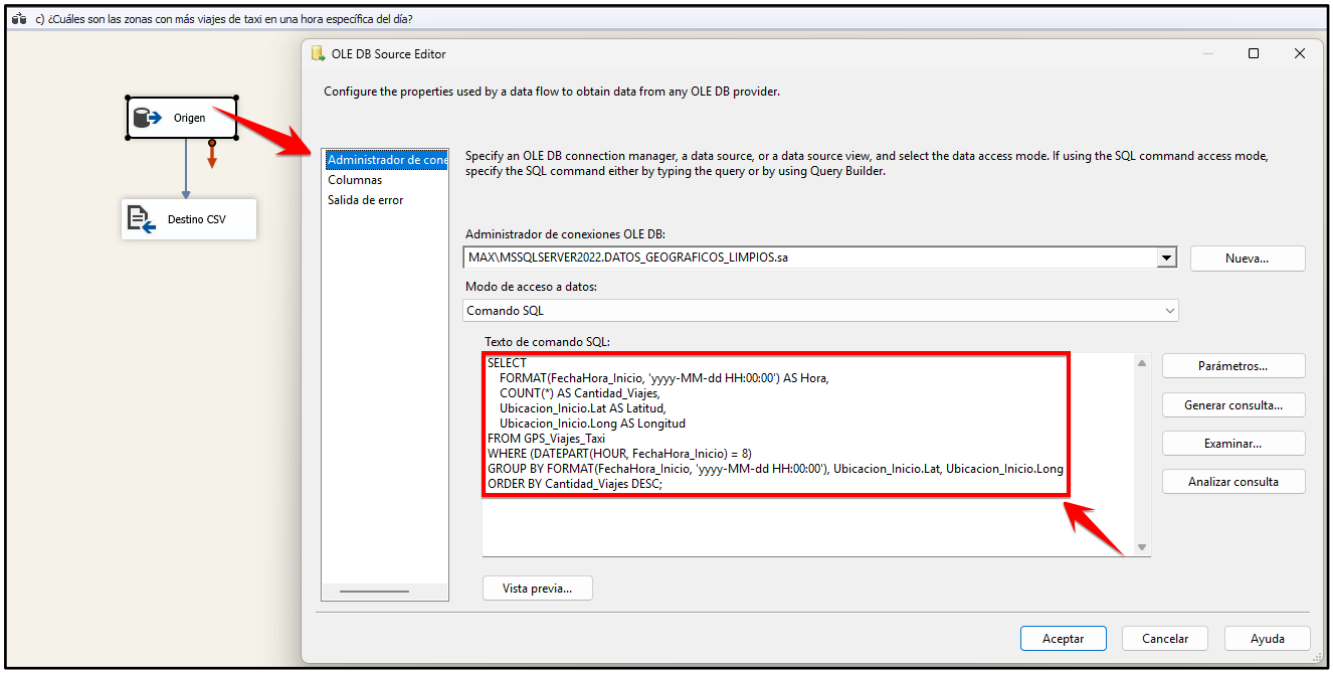


Destino: Un destino CSV.

c) ¿Cuáles son las zonas con más viajes de taxi en una hora específica del día?

Objetivo: Identificar las zonas con mayor concentración de viajes en un horario específico, en este caso a las 8:00.

Extrayendo datos en SSIS mediante consulta SQL:

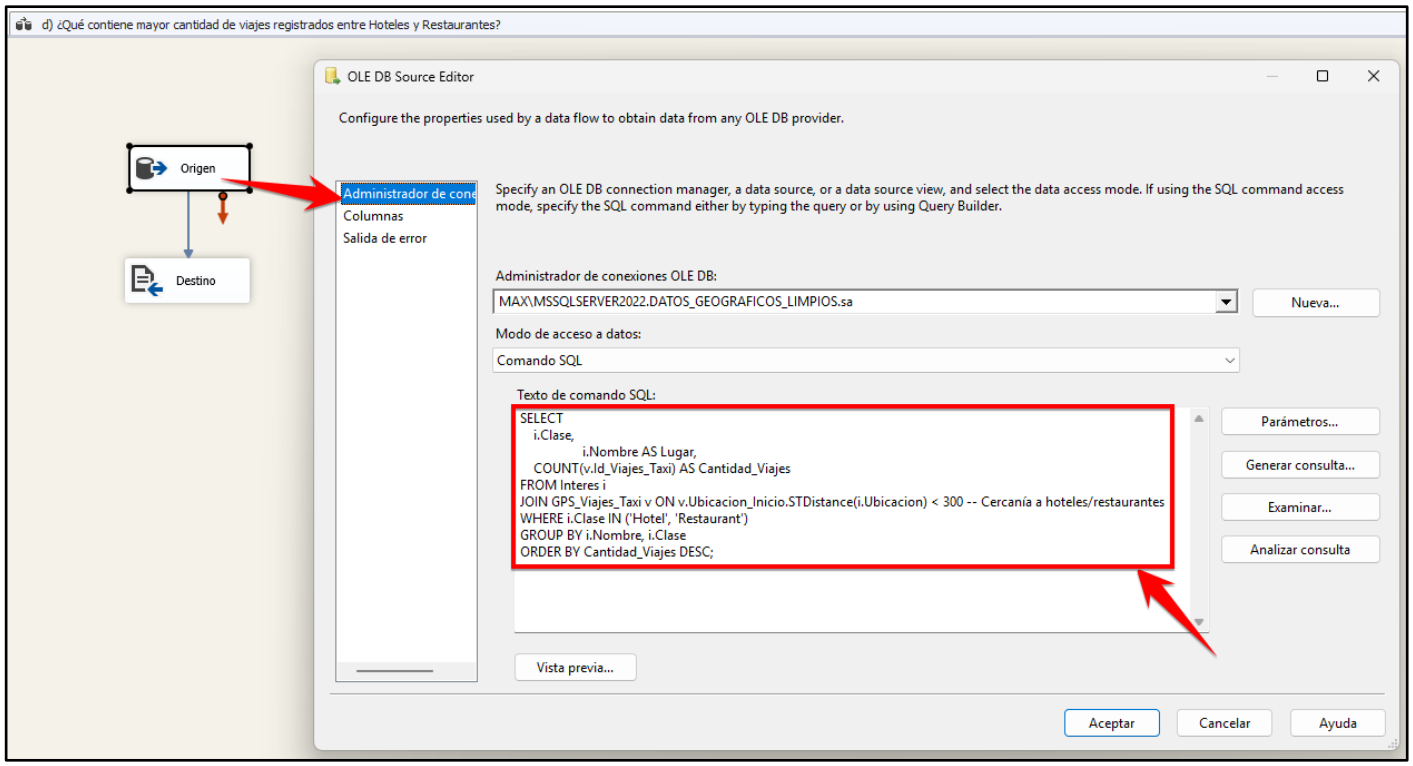


Destino: Un destino CSV.

d) ¿Qué punto de interés contiene mayor cantidad de viajes registrados entre Hoteles y Restaurantes?

Objetivo: Extraer cantidad de viajes entre Hoteles y Restaurantes e identificar cual tiene más.

Extrayendo datos en SSIS mediante consulta SQL:

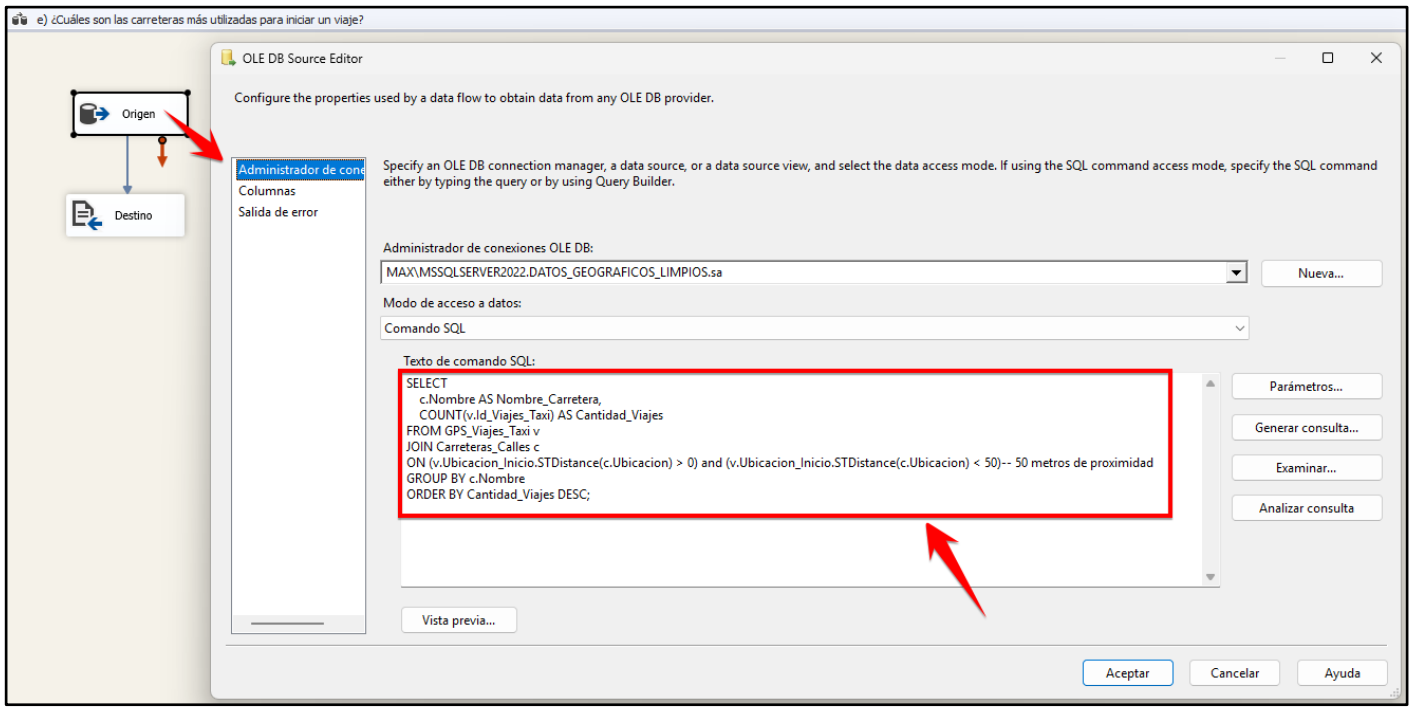


Destino: Un destino CSV.

e) ¿Cuáles son las carreteras más utilizadas para iniciar un viaje?

Objetivo: Determinar en qué carreteras inician más viajes.

Extrayendo datos en SSIS mediante consulta SQL:

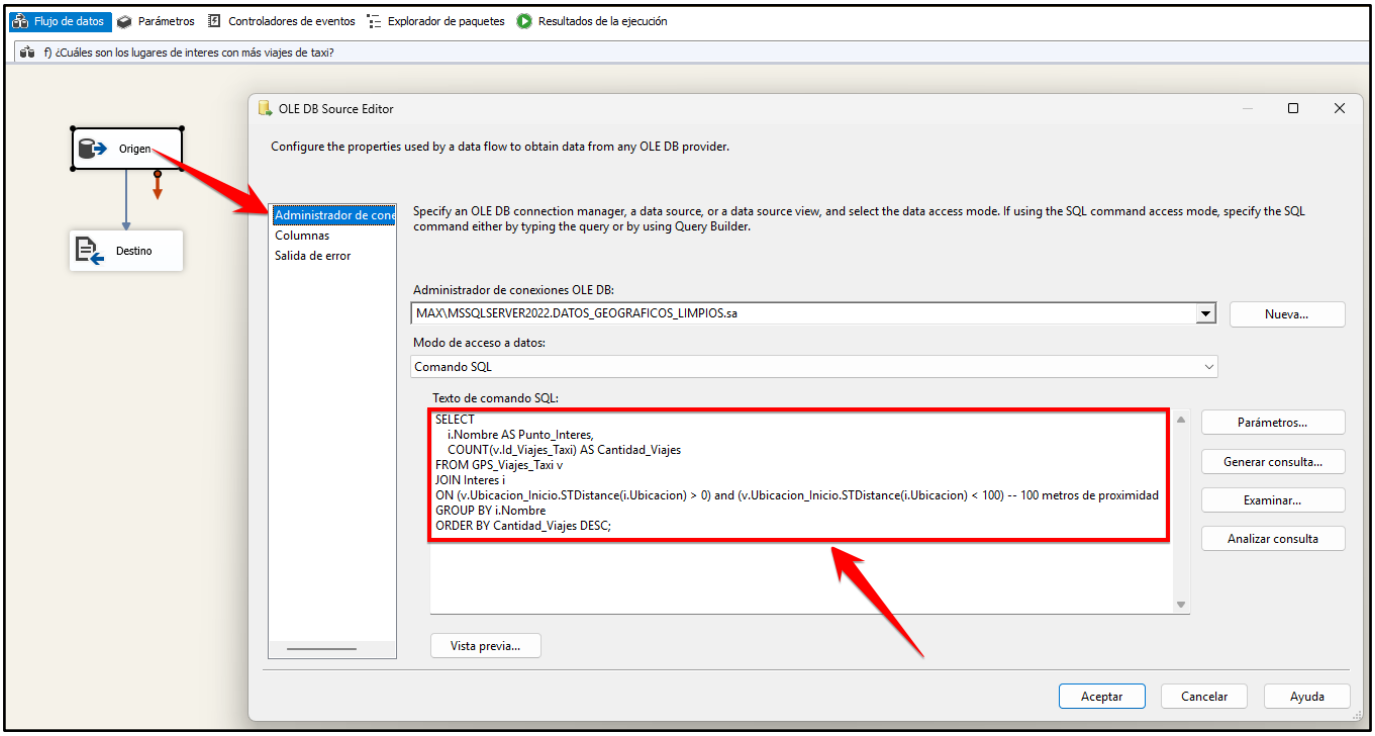


Destino: Un destino CSV.

f) ¿Cuáles son los lugares de interés con más viajes de taxi?

Objetivo: Determinar qué lugares de interés generan más tráfico de taxis.

Extrayendo datos en SSIS mediante consulta SQL:

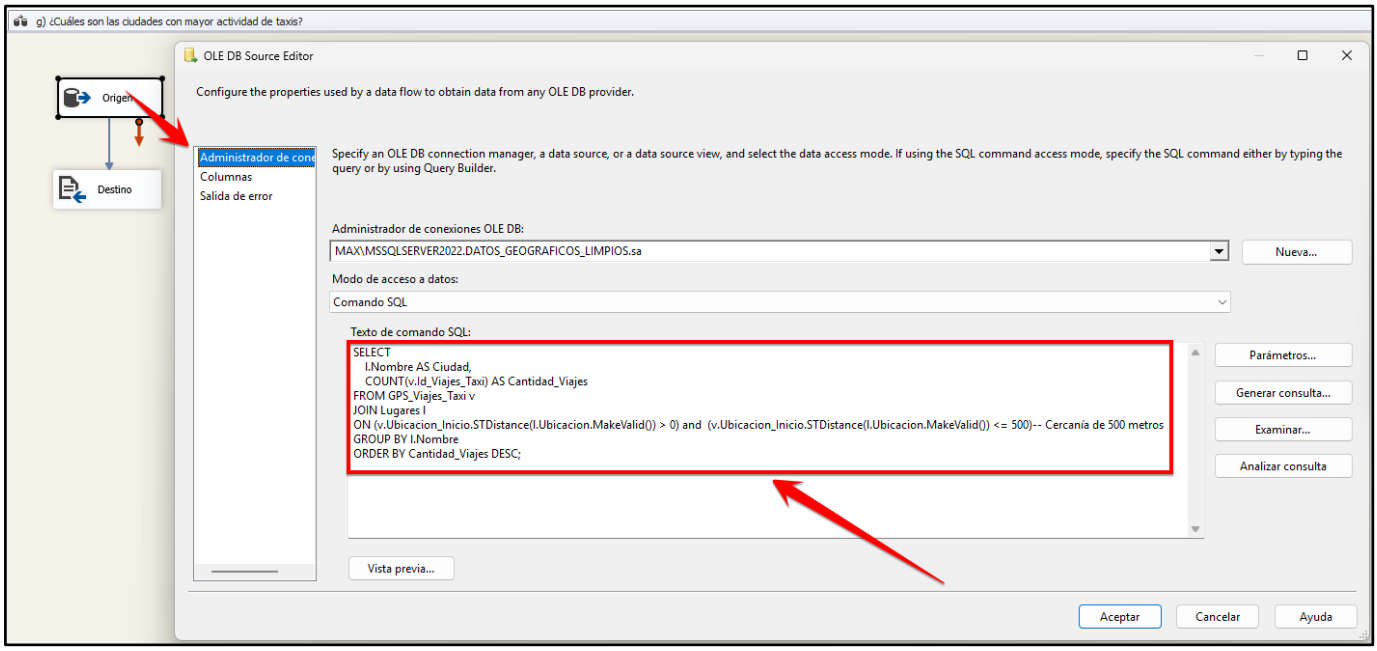


Destino: Un destino CSV.

g) ¿Cuáles son las ciudades con mayor actividad de taxis?

Objetivo: Identificar en qué ciudades hay más viajes.

Extrayendo datos en SSIS mediante consulta SQL:

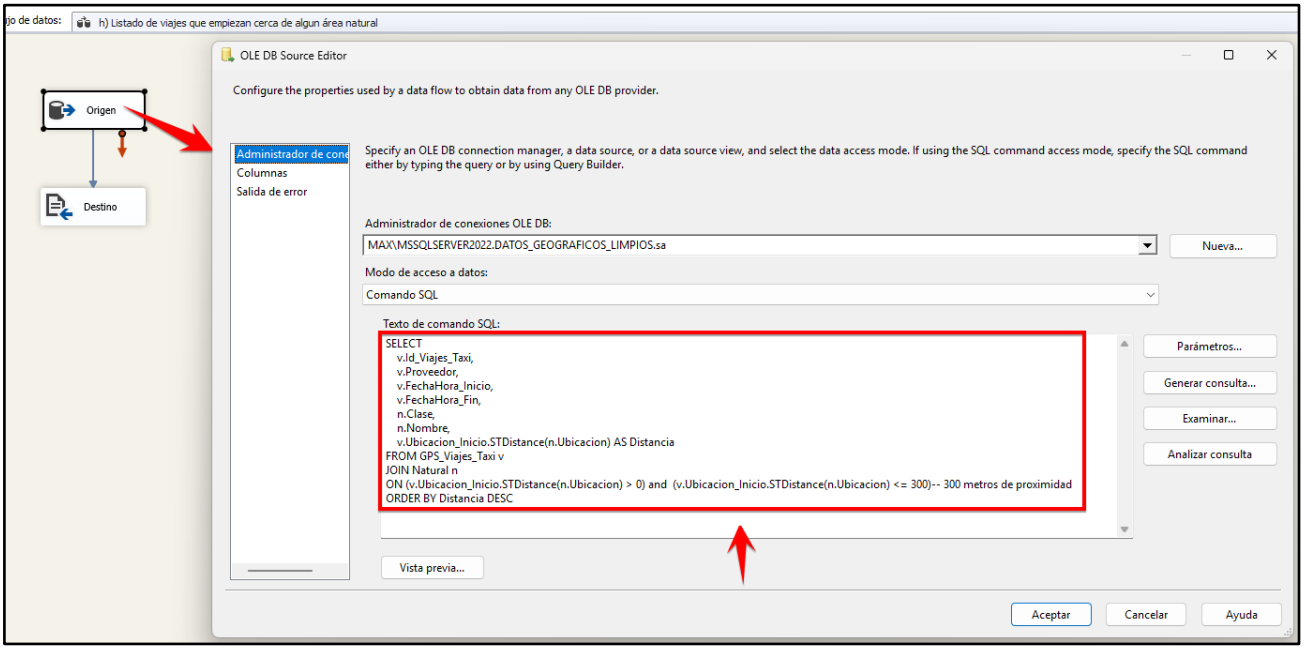


Destino: Un destino CSV.

h) Listado de viajes que empiezan cerca de algún área natural.

Objetivo: Identificar viajes que inician en un área natural.

Extrayendo datos en SSIS mediante consulta SQL:

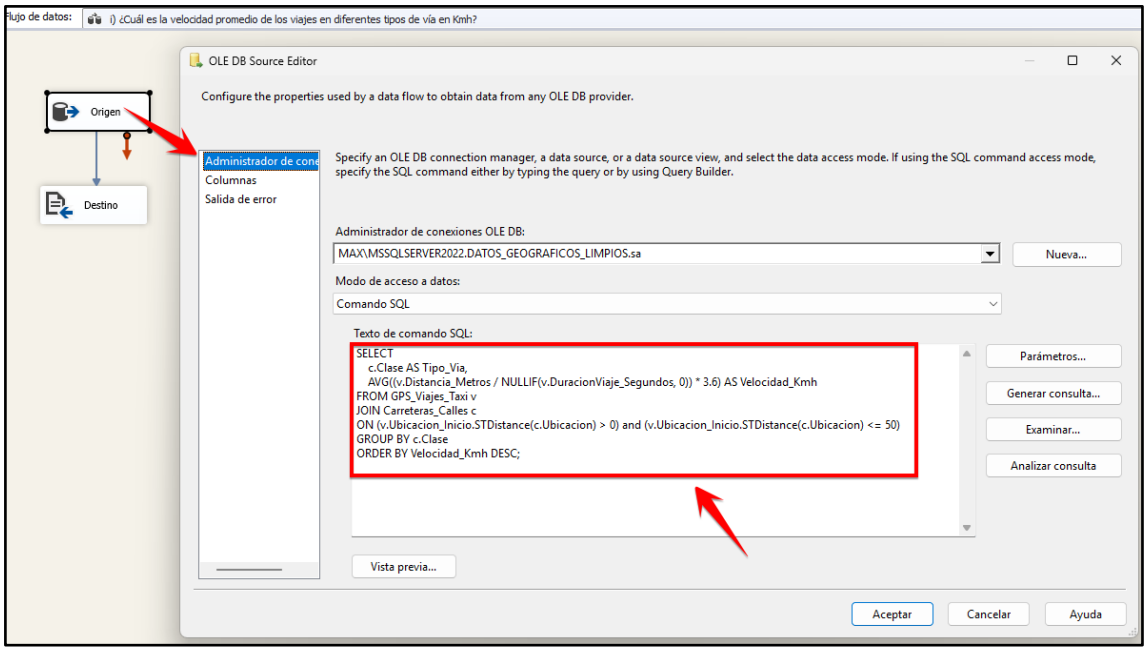


Destino: Un destino CSV.

i) ¿Cuál es la velocidad promedio de los viajes en diferentes tipos de vía en Km/h?

Objetivo: Identificar viajes que inician en un área natural.

Extrayendo datos en SSIS mediante consulta SQL:

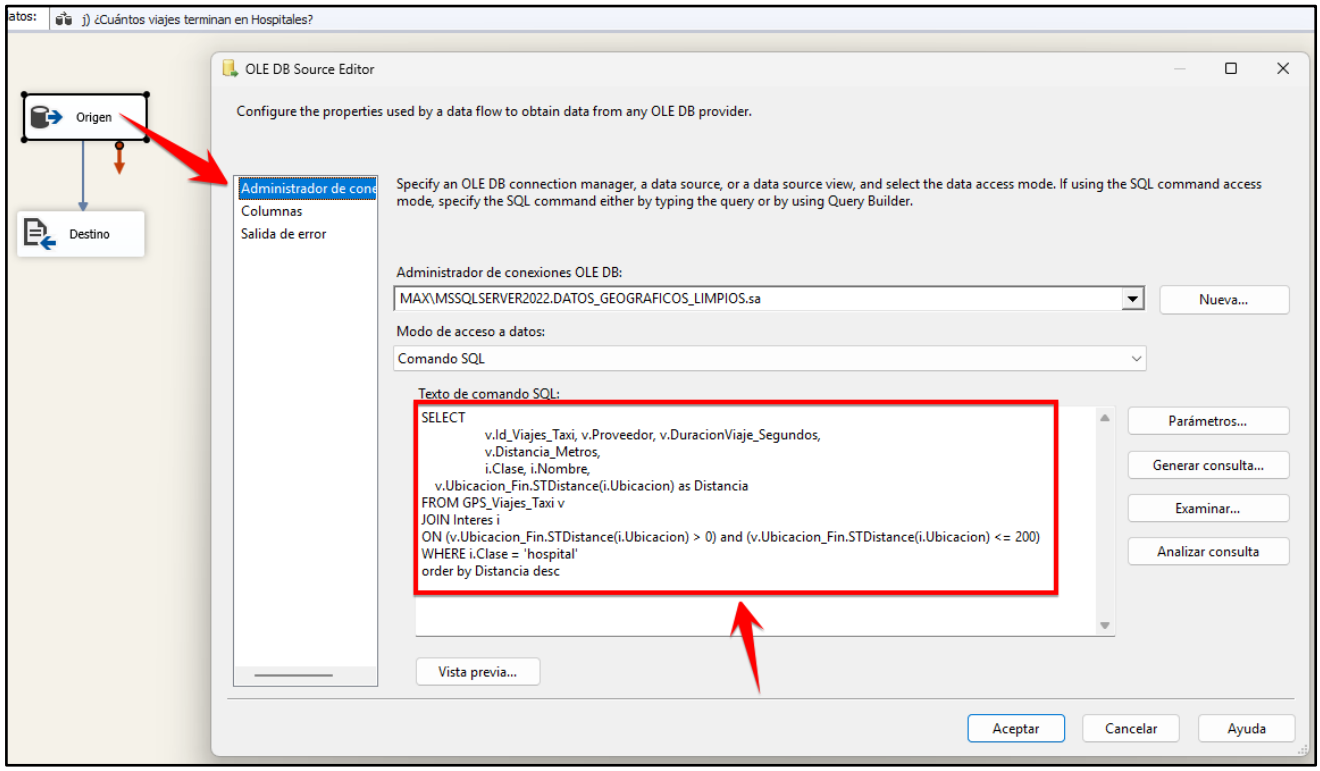


Destino: Un destino CSV.

j) ¿Cuántos viajes terminan en Hospitales?

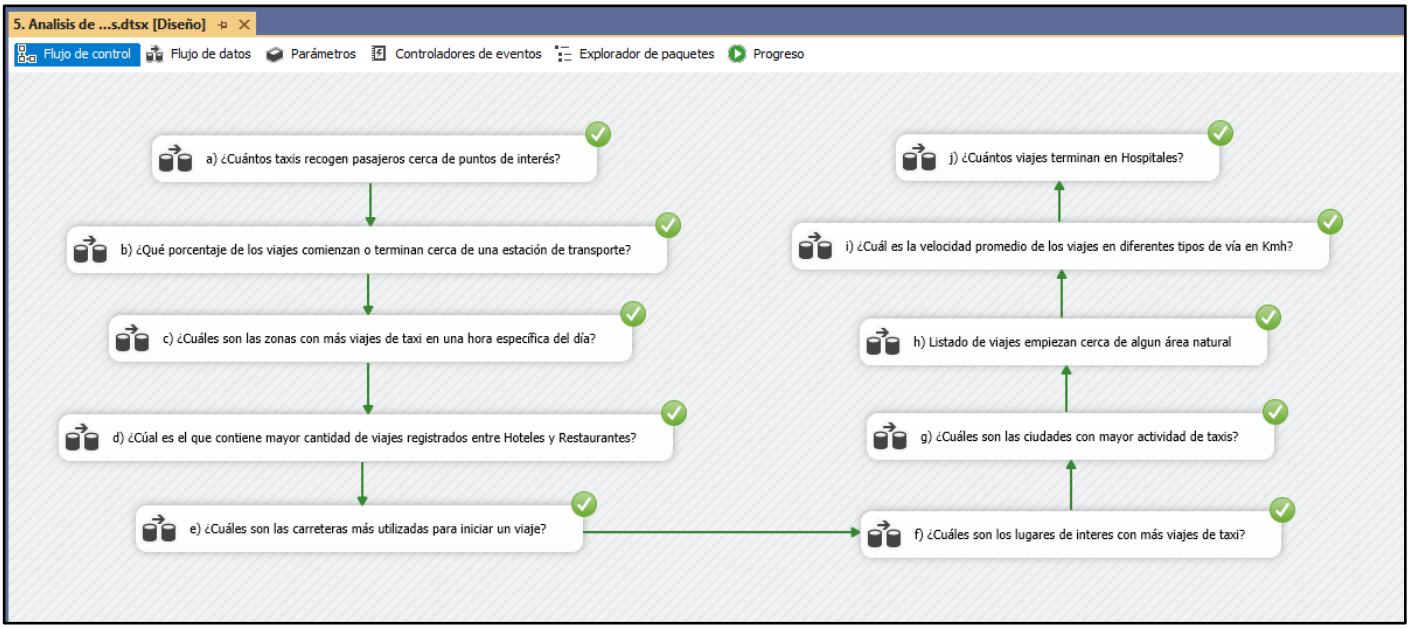
Objetivo: Medir la actividad de taxis en áreas de hospitales.

Extrayendo datos en SSIS mediante consulta SQL:



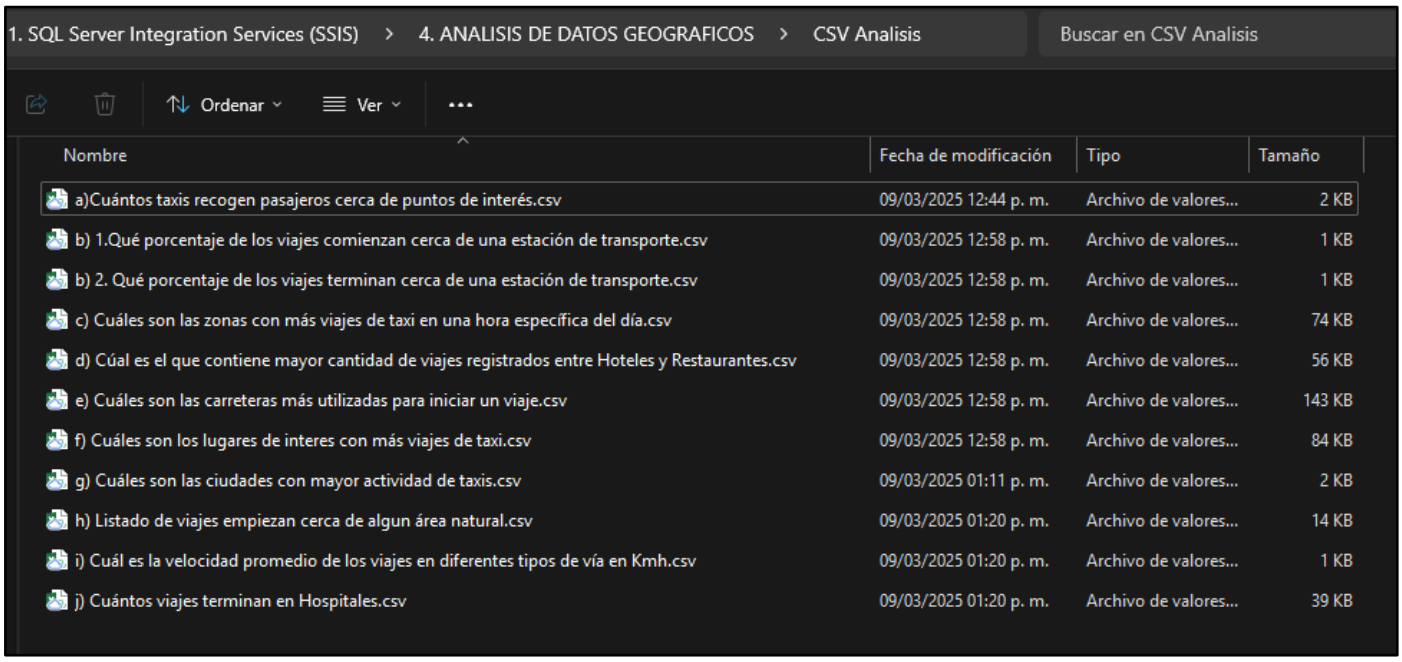
Destino: Un destino CSV.

EJECUTANDO LAS TAREAS DE FLUJO:



III. REVISIÓN DE ANÁLISIS

Al ejecutar el proyecto se observa que cada CSV se guardó correctamente. Y revisaremos cada uno de ellos:



1. **Analizando Resultados:**

   a) **¿Cuántos taxis recogen pasajeros cerca de puntos de interés (hoteles, restaurantes, estaciones de transporte, parques, ¿etc.)?**

   Resultado CSV:

	A	B
1	Nombre_Clase	Cantidad_Viajes
2	restaurant	4427
3	convenience	4407
4	fast_food	3651
5	bank	3643
6	cafe	3147
7	supermarket	2455

Conclusión: El punto de Interés de “Restaurant” tiene un total de 4427 viajes de taxi, seguido de “Convenience” con 4407 viajes, y “fast_food” con 3651, estos son los puntos de interés que más viajes tienen.

b) ¿Qué porcentaje de los viajes comienzan o terminan cerca de una estación de transporte?

Resultado CSV:

A
Porcentaje_Cerca
41.643324

A
Porcentaje_Cerca
38.05788982

Conclusión: El 41.64% de los viajes iniciaron cerca de una estación de transporte y el 38.05% terminaron cerca de una.

c) ¿Cuáles son las zonas con más viajes de taxi en una hora específica del día?

Resultado CSV:

A	B	C	D
Hora	Cantidad_Viajes	Latitud	Longitud
20/01/2017 08:00	3	19.35577939	-99.06293764
13/07/2016 08:00	2	19.53285743	-99.02608909
23/11/2016 08:00	2	19.43853029	-99.17956287
08/12/2016 08:00	2	19.2352431	-99.09838262
02/06/2017 08:00	2	19.47776544	-99.09410276
19/11/2016 08:00	2	19.3309183	-99.0698295
18/05/2017 08:00	2	19.6034116	-99.0278804
25/05/2017 08:00	2	19.26768838	-99.21132346

Conclusión: La zona con latitud y longitud mostrada en la imagen representan la cantidad de viajes que se realizaron en esa hora, en este caso, 3 viajes en la misma zona, el mismo día, en el horario de las 8:00 a.m.

d) ¿Qué punto de interés contiene mayor cantidad de viajes registrados entre Hoteles y Restaurantes?

Resultado CSV:

A	B	C
Clase	Lugar	Cantidad_Viajes
restaurant	Vips	741
restaurant	Los Arcos	678
restaurant	La Casa de Toño	639
restaurant	Casa de Pepe	638
restaurant	Cambalache	632
restaurant	Sanborns	629
restaurant	Gino's Insurgentes	626
restaurant	Munchies	613
restaurant	Chilli's	612
restaurant	Cortes Recreo	571

Conclusión: Se observa que el que contiene mayor cantidad de viajes son realizados por la Clase Restaurant.

e) ¿Cuáles son las carreteras más utilizadas para iniciar un viaje?

Resultado CSV:

A	B
Nombre_Carretera	Cantidad_Viajes
Avenida Insurgentes Sur	2614
Calle Lago Alberto	890
Calle Lago Xochimilco	871
Prolongación Lago Tana	772
Avenida Morelos	430

Conclusión: La Avenida Insurgentes Sur fue la vía más utilizada como punto de partida.

f) ¿Cuáles son los lugares de interés con más viajes de taxi?

Resultado CSV:

A	B
Punto_Interes	Cantidad_Viajes
Oxxo	977
7-Eleven	562
BBVA	477
HSBC	449
La Casa de Las Enchiladas (Lago Alberto)	447
Inbursa	446
Olivo	446

Conclusión: El Oxxo resultó ser uno de los lugares de interés con más actividad de taxis.

g) ¿Cuáles son las ciudades con mayor actividad de taxis?

Resultado CSV:

A	B
Ciudad	Cantidad_Viajes
Ciudad de México	257
La Condesa	89
Pedregal de Tepepan	67
La Roma	24

Conclusión: La Ciudad de México concentró la mayoría de los inicios de viaje.

h) Listado de viajes que empiezan cerca de algún área natural.

Resultado CSV:

A	B	C	D	E	F	G
d_Viajes_Taxi	Proveedor	FechaHora_Inicio	FechaHora_Fin	Clase	Nombre	Distancia
1742	Mexico DF Taxi de Sitio	26/11/2016 03:16	26/11/2016 04:18	tree	Árbol de la Noche Victoriosa	299.7353015
9915	Mexico DF Taxi de Sitio	09/07/2017 04:53	09/07/2017 04:59	tree	Palmera	299.0344201
9912	Mexico DF Taxi de Sitio	09/07/2017 02:10	09/07/2017 02:32	spring	La fuente de Liverpool	297.9226102
3674	Mexico DF Taxi de Sitio	01/12/2016 10:30	01/12/2016 12:05	tree	Trueno	297.5944573
3526	Mexico DF Taxi de Sitio	16/10/2016 12:07	16/10/2016 12:09	tree	Ahuehuete El Sargento	296.6772882
5198	Mexico DF Taxi de Sitio	16/11/2016 05:14	16/11/2016 06:19	tree	Palmera	294.5908749
3674	Mexico DF Taxi de Sitio	01/12/2016 10:30	01/12/2016 12:05	tree	Trueno	292.3881281
9159	Mexico DF Taxi de Sitio	28/06/2017 12:12	28/06/2017 12:38	peak	Cerro de Chapultepec	288.0380903
7415	Mexico DF Taxi de Sitio	27/05/2017 02:44	27/05/2017 03:05	tree	Palmera	285.5610089
10485	Mexico DF Taxi de Sitio	10/04/2017 08:05	10/04/2017 09:48	tree	El Cardenal	285.1579763
3283	Mexico DF Taxi Libre	12/05/2017 11:41	12/05/2017 11:50	tree	Antiguo Ahuehuete. Monumento de Tacuba	283.0257782
149	Mexico DF Taxi Libre	22/04/2017 09:54	22/04/2017 10:04	tree	El Cardenal	282.6644205
3674	Mexico DF Taxi de Sitio	01/12/2016 10:30	01/12/2016 12:05	tree	Trueno	280.4082849
508	Mexico DF Taxi de Sitio	01/04/2017 03:07	01/04/2017 03:56	tree	Trueno	278.8156672

Conclusión: Se identificaron múltiples viajes que iniciaron cerca de zonas naturales, lo cual podría usarse para evaluar la demanda de transporte en áreas recreativas o rurales.

i) ¿Cuál es la velocidad promedio de los viajes en diferentes tipos de vía en Km/h?

Resultado CSV:

A	B
Tipo_Via	Velocidad_Kmh
trunk_link	42.48
living_street	19.16648
motorway_link	18.211764
motorway	18.077182
busway	18
cycleway	17.571428
secondary_link	17.485714
trunk	17.37348
primary_link	17.2
primary	16.892484
pedestrian	16.438554
unclassified	16.253465
footway	15.688235
service	15.651752
path	15.463636
secondary	15.353791
residential	15.013888
tertiary	14.592934
steps	11.59266

Conclusión: Se obtuvieron velocidades promedio por tipo de vía, útiles para análisis de tráfico.

j) ¿Cuántos viajes terminan en Hospitales?

Resultado CSV:

Id_Viajes_Taxi	Proveedor	DuracionViaje_Segundos	Distancia_Metros	Clase	Nombre	Distancia
8479	Mexico DF Taxi Libre	4062	16026	hospital	Hospital Pediatrico Legaria	199.9852634
9300	Mexico DF Taxi Libre	1208	5540	hospital	Hospital Santa Monica	199.679417
1648	Mexico DF Taxi de Sitio	1567	5010	hospital	Médica San Luis	199.4878467
1865	Mexico DF Taxi Libre	649	5669	hospital	Centro de Salud Dr. D. Orvañanos	199.4445984
9565	Mexico DF Taxi Libre	473	3623	hospital	ISSSTE Clinica de Medicina Familiar	199.2290968
4999	Mexico DF Taxi de Sitio	1377	3039	hospital	Ortopedia Flores	198.9740155
2036	Mexico DF Taxi de Sitio	1869	9683	hospital	Hospital Santa Elena, Angeles Roma	198.9570082
5740	Mexico DF Taxi Libre	823	7527	hospital	Hospital Materno Infantil Dr. Nicolas M. Cedillo	198.9114922
4184	Mexico DF Taxi Libre	253	1846	hospital	Centro de Salud Cardiel	198.3264823
1064	Mexico DF Taxi Libre	2783	8786	hospital	Santa Coleta	198.1805927
2813	Mexico DF Taxi de Sitio	915	4334	hospital	Hospital Pediatrico Legaria	197.989997
7650	Mexico DF Taxi de Sitio	954	4263	hospital	Hospital Pediatrico Legaria	197.7969382
4705	Mexico DF Radio Taxi	1170	35412	hospital	Imss Villalonguin	197.7563328
8916	Mexico DF Radio Taxi	1287	11654	hospital	Centro de Salud Cardiel	197.6331103
2912	Mexico DF Taxi Libre	1704	7194	hospital	Clinica Imss	197.487134
8834	Mexico DF Taxi Libre	211	910	hospital	Hospital Boutique	197.3746734
7487	Mexico DF Taxi Libre	838	2680	hospital	Centro de Salud Cardiel	196.9284487

Conclusión: Muestra un listado de más de 400 registros de viajes de taxis que tienen como destino final cercano un Hospital.

IV. CONCLUSIONES

A lo largo del desarrollo de este proyecto, se logró implementar con éxito un flujo completo de integración, transformación y análisis de datos geográficos utilizando herramientas como SSIS, SQL Server, QGIS y lenguajes como C#. El objetivo principal fue aprovechar datos espaciales (Shapefiles) y datos de GPS de viajes en taxi para obtener información valiosa sobre el comportamiento del transporte urbano en la Ciudad de México.

Entre los principales logros se destacan:

- **Integración efectiva de datos geoespaciales** en un entorno de base de datos relacional, superando los retos que implica el manejo de datos de tipo GEOGRAPHY y GEOMETRY.
- **Depuración y filtrado de datos**, lo que permitió asegurar la calidad de la información utilizada en los análisis.
- **Automatización del proceso ETL**, facilitando futuras actualizaciones o replicación del proyecto en otras ciudades.
- **Obtención de indicadores clave**, como zonas con mayor demanda de taxis, puntos de interés con mayor actividad, uso de carreteras, y comportamiento según tipo de vía.
- **Visualización y exportación de resultados** en archivos CSV, útiles para informes ejecutivos o como insumos para otras herramientas analíticas o visualización de datos.

Este proyecto me permitió darme cuenta de cómo el uso de datos espaciales junto con herramientas de análisis puede aportar información muy valiosa para tomar decisiones en temas como el transporte urbano, la planificación territorial o incluso para entender mejor cómo se mueve la ciudad y cómo se podría mejorar la movilidad.

Este proyecto fue desarrollado como parte de mi portafolio profesional en análisis de datos geográficos.