



Datawarehouse y Minería de Datos

Docente: Karens Medrano

Estudiante: Eduardo Ezequiel López Rivera

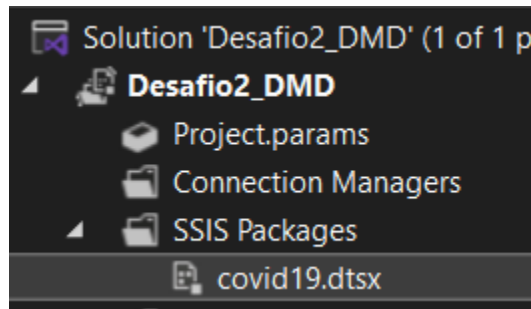
Carné: LR230061

Ciclo: 01-2025

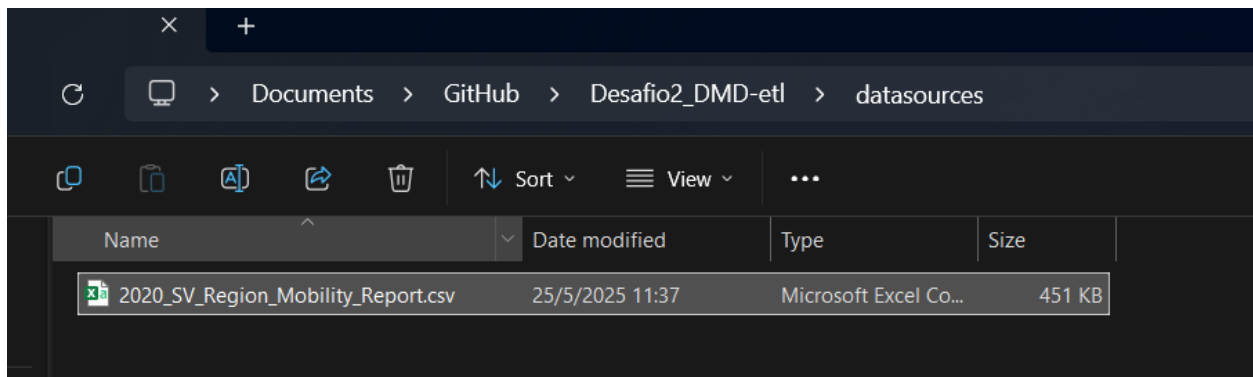
Desafío #2

Creación de proceso ETL

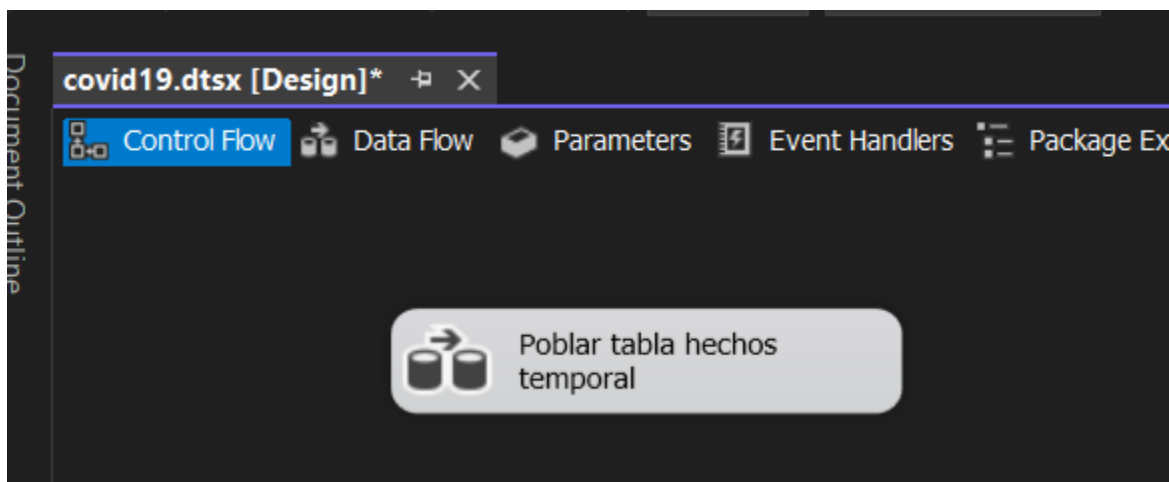
Creamos un proyecto en Visual Studio 2022 de tipo Integration Services Project, renombramos nuestro paquete SSIS a covid19.dtsx



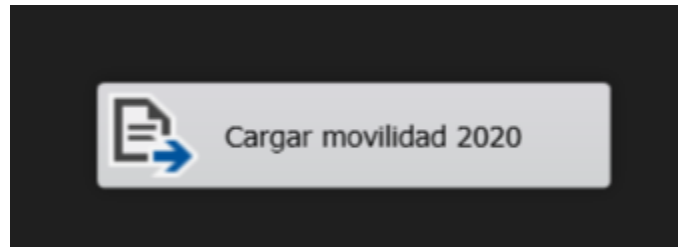
Descargamos el informe de movilidad de COVID-19 de El Salvador del año 2020, obtenemos el siguiente archivo en formato .CSV



En nuestro Control Flow, creamos un Data Flow Task, el cual nos servirá para ordenar la data del CSV, la nombramos “Poblar tabla hechos temporal”



Agregamos un Flat File Source, el cual llamaremos “Cargar movilidad 2020”.



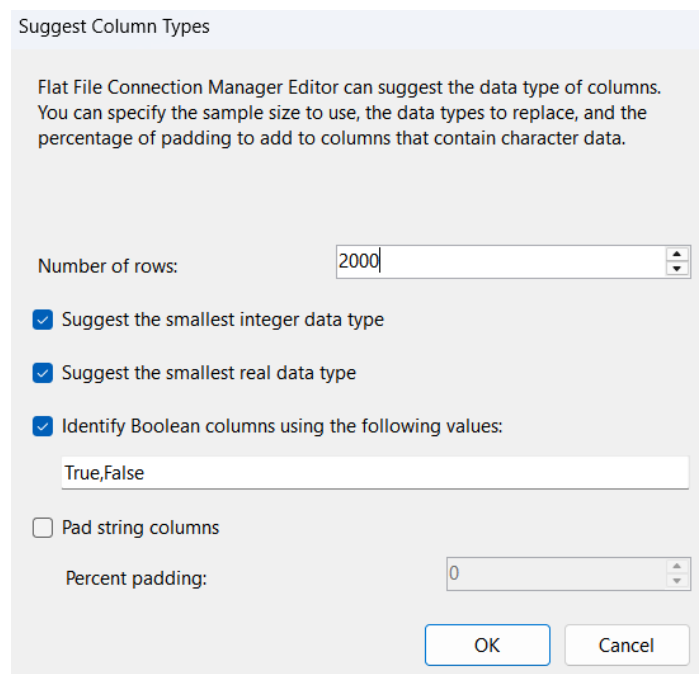
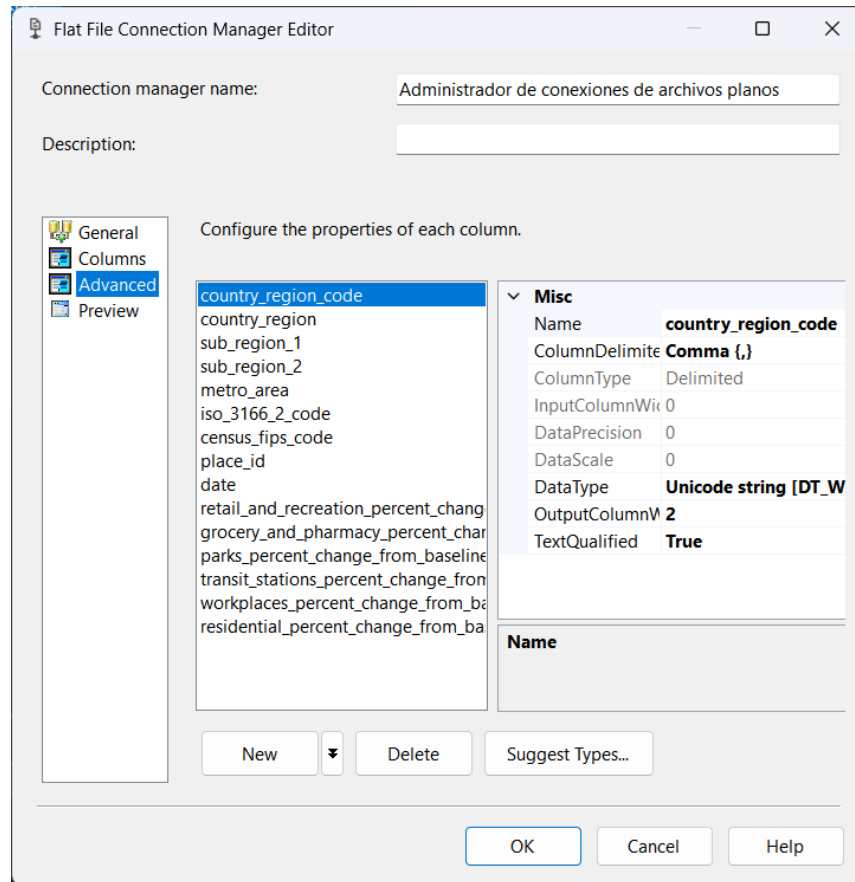
Configuramos el Flat File Source con los siguientes parámetros (Ruta del archivo .CSV, Code page en 65001 UTF-8 para que aparezcan las tildes de forma correcta)

The image shows a screenshot of the "Flat File Connection Manager Editor" window. The window has a title bar with standard Windows controls. On the left is a sidebar with four tabs: "General" (selected), "Columns", "Advanced", and "Preview". The main area is titled "Select a file and specify the file properties and the file format." and contains the following fields:

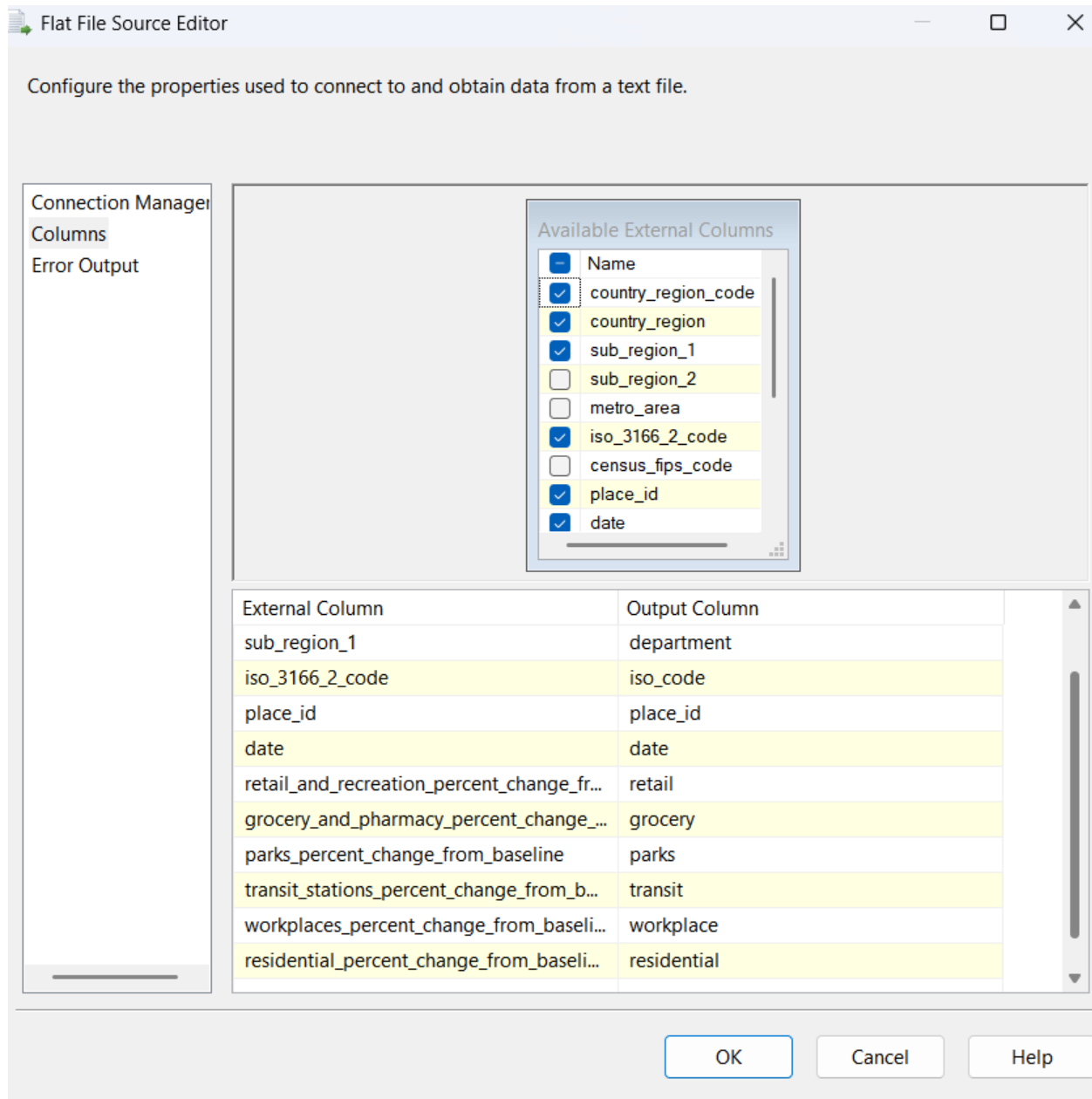
- "Connection manager name:" with the text "Administrador de conexiones de archivos planos".
- "Description:" with an empty text box.
- "File name:" with the text "\\2020_SV_Region_Mobility_Report.csv" and a "Browse..." button.
- "Locale:" with a dropdown menu showing "Spanish (El Salvador)" and an unchecked "Unicode" checkbox.
- "Code page:" with a dropdown menu showing "65001 (UTF-8)".
- "Format:" with a dropdown menu showing "Delimited".
- "Text qualifier:" with a text box containing "<none>".
- "Header row delimiter:" with a dropdown menu showing "{CR}{LF}".
- "Header rows to skip:" with a spinner box set to "0".
- A checked checkbox labeled "Column names in the first data row".

At the bottom right are three buttons: "OK", "Cancel", and "Help".

Configuramos los tipos de datos en la pestaña Columns, todos los strings los convertimos a Unicode string y utilizando la herramienta “Suggest Types” obtenemos un OutputColumnWidth para cada campo y un tipado de datos sugerido para los demás campos, calculado de los primeros 2000 registros del .CSV



En el Flat File Source editor nos deshacemos de algunas columnas que vienen vacías porque no aplican para nuestro país, así mismo, cambiamos de nombre algunas columnas para mayor simplicidad.



Conectamos un “Unpivot” para transformar las columnas que tienen datos de movilidad en una sola y sus respectivos valores en otra.

Unpivot Transformation Editor

Specify the columns to pivot into rows to make an unnormalized dataset into a more normalized version.

Available Input Columns

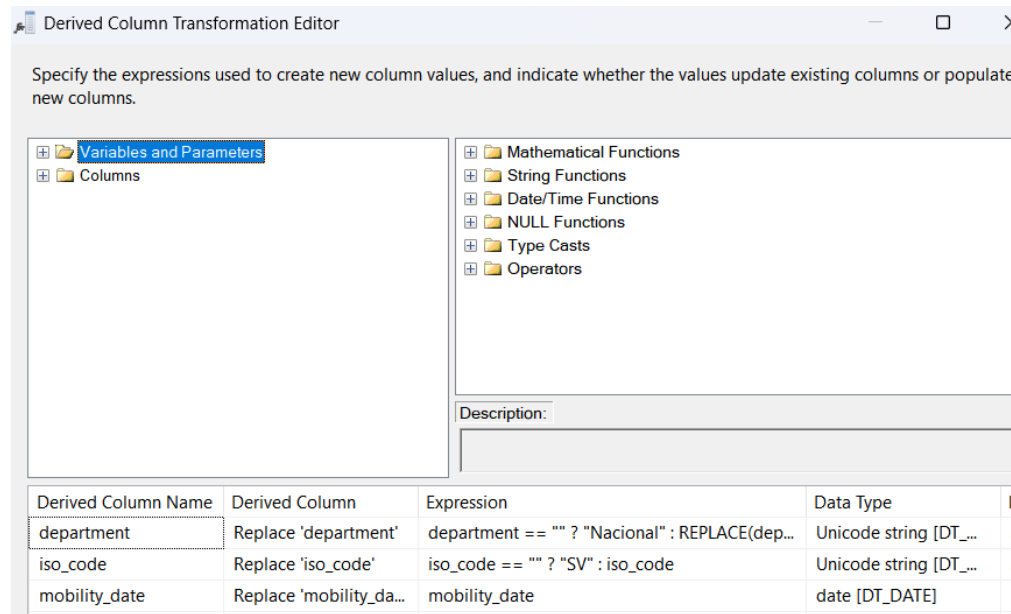
<input type="checkbox"/>	Name	Pass Thr...
<input checked="" type="checkbox"/>	retail	<input type="checkbox"/>
<input checked="" type="checkbox"/>	grocery	<input type="checkbox"/>
<input checked="" type="checkbox"/>	parks	<input type="checkbox"/>
<input checked="" type="checkbox"/>	transit	<input type="checkbox"/>
<input checked="" type="checkbox"/>	workplace	<input type="checkbox"/>
<input checked="" type="checkbox"/>	residential	<input type="checkbox"/>

Input Column	Destination Column	Pivot Key Value
retail	porcentaje_cambio	retail
grocery	porcentaje_cambio	grocery
parks	porcentaje_cambio	parks
transit	porcentaje_cambio	transit
workplace	porcentaje_cambio	workplace
residential	porcentaje_cambio	residential

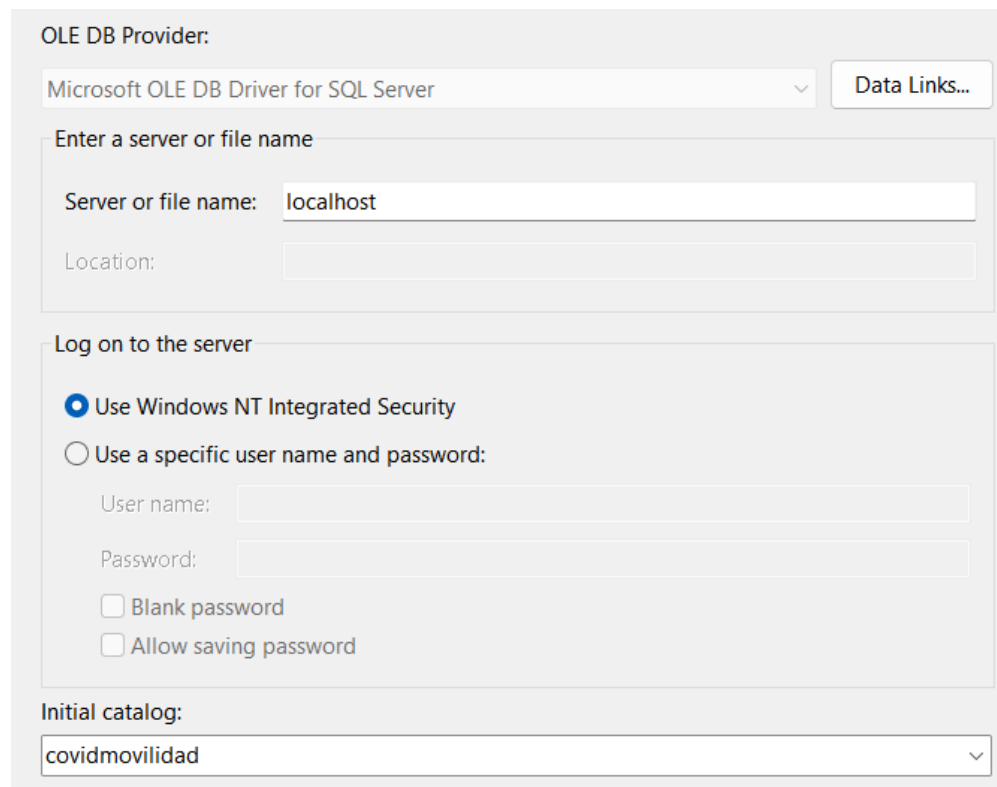
Pivot key value column name:

tipo_movilidad

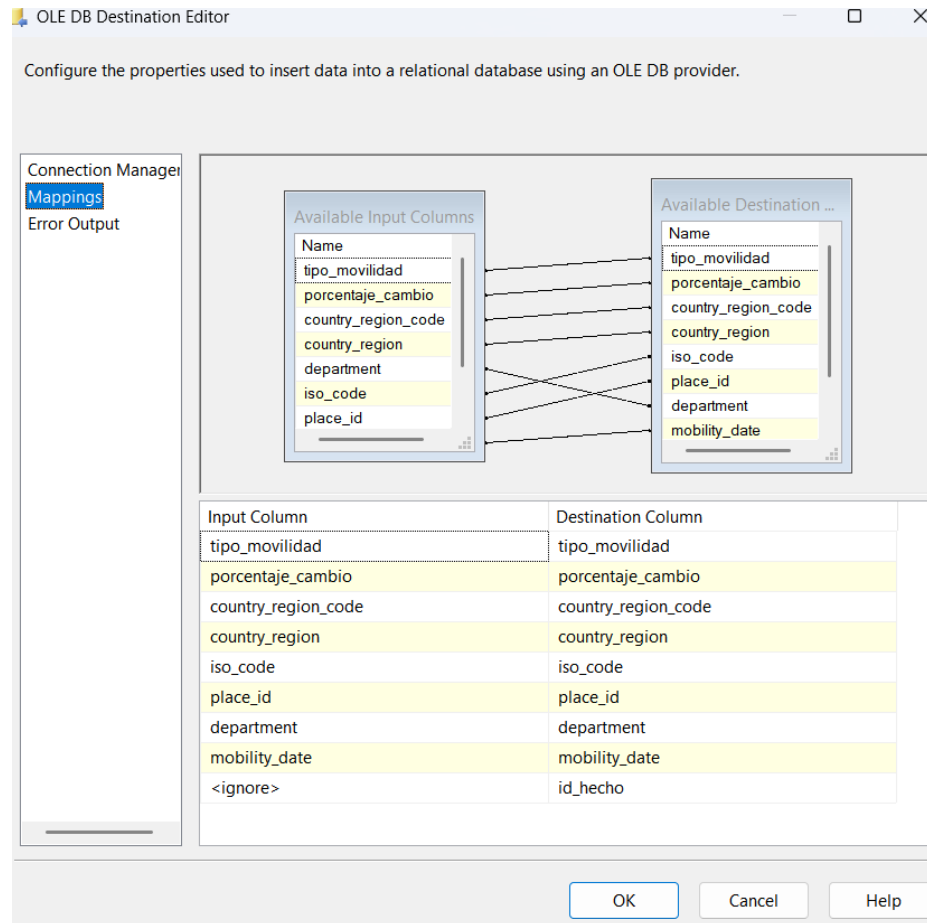
Conectamos un “Derived column” para limpiar los datos de la columna Department, poblar los datos vacíos en iso_code, y reemplazamos date por mobility_date.



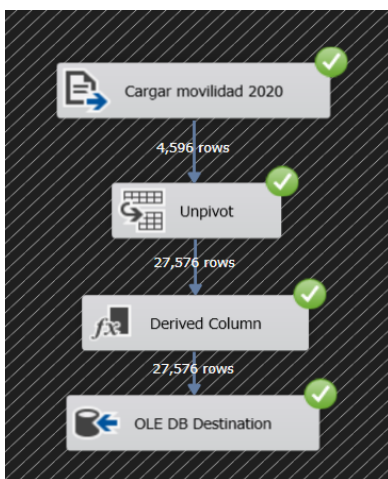
Finalmente agregamos un OLE DB Destination para guardar la data en nuestra base de datos en SQL Server. Configuramos la conexión a la base de datos.



Creamos la tabla “hechos” en nuestra base de datos y enlazamos las columnas correspondientes

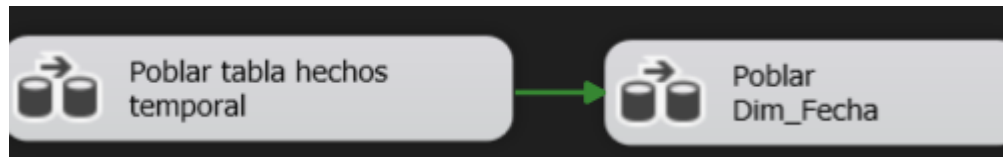


Ejecutamos con éxito el flujo.



Results									
	tipo_movilidad	porcentaje_cambio	country_region_code	country_region	department	iso_code	place_id	mobility_date	id_hecho
1	grocery	5	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	1
2	parks	0	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	2
3	residential	-1	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	3
4	retail	4	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	4
5	transit	-1	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	5
6	workplace	4	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-15 00:00:00.000	6
7	grocery	6	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	7
8	parks	1	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	8
9	residential	0	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	9
10	retail	4	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	10
11	transit	1	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	11
12	workplace	0	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-16 00:00:00.000	12
13	grocery	4	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-17 00:00:00.000	13
14	parks	-3	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-17 00:00:00.000	14
15	residential	0	SV	El Salvador	Nacional	SV	ChIJVwZkWaYnY48RMkiFmOsWmm8	2020-02-17 00:00:00.000	15

Procederemos a crear 3 dimensiones a partir de nuestra tabla de hechos temporal, comenzamos con Dim_Fecha, creamos otro Data Flow Task, le llamaremos “Poblar Dim_Fecha”



Colocamos un OLE DB Source dentro del flujo y obtenemos las fechas de los registros sin repetir con la siguiente consulta SQL y les damos un formato segmentado.

OLE DB Source Editor

Configure the properties used by a data flow to obtain data from any OLE DB provider.

Connection Manager
Columns
Error Output

Specify an OLE DB connection manager, a data source, or a data source view, and select the data access mode. If using the SQL command access mode, specify the SQL command either by typing the query or by using Query Builder.

OLE DB connection manager:
localhost.covidmovilidad

Data access mode:
SQL command

SQL command text:

```
SELECT  
DISTINCT  
mobility_date,  
YEAR(mobility_date) AS year_data,  
MONTH(mobility_date) AS month_data,  
DAY(mobility_date) AS day_data,  
DATENAME(MONTH, mobility_date) AS month_name,  
DATENAME(WEEKDAY, mobility_date) AS day_month,  
CONVERT(CHAR(8), mobility_date, 112) AS id_date  
FROM hechos  
WHERE mobility_date IS NOT NULL
```

Parameters...
Build Query...
Browse...
Parse Query

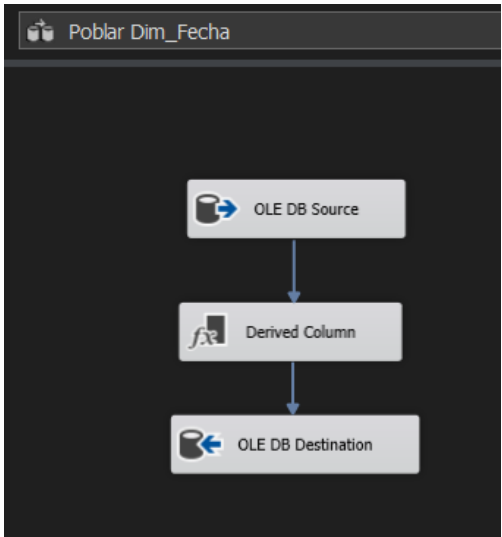
Preview...

OK Cancel Help

Convertimos la columna id_date recién creada a integer

Derived Column Name	Derived Column	Expression
id_date	Replace 'id_date'	(DT_I4)TRIM(id_date)

Conectamos un OLE DB Destination y creamos la tabla Dim_Fecha, mapeamos las columnas.



Configure the properties used to insert data into a relational database using an OLE DB provider.

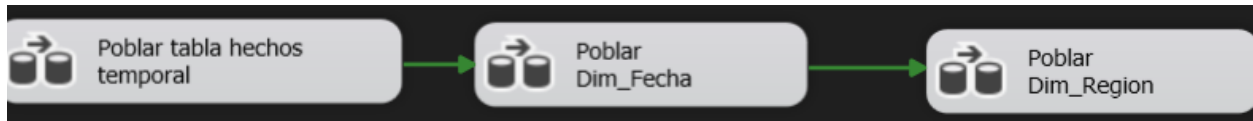
Connection Manager
Mappings
Error Output

Available Input ...

Available Desti...

Input Column	Destination Column
mobility_date	mobility_date
year_data	year_data
month_data	month_data
day_data	day_data
month_name	month_name
day_month	day_month
id_date	id_date

Creamos otro flujo para poblar Dim_Region.



Conservamos una estructura muy simple, obtenemos datos mediante una consulta, en la misma consulta cambiamos un poco las columnas, y al crear la tabla Dim_Region, creamos un campo id_region primary key.

Data access mode:

SQL command

SQL command text:

```
SELECT DISTINCT
country_region,
department
FROM hechos
WHERE department IS NOT NULL
```

Parameters...

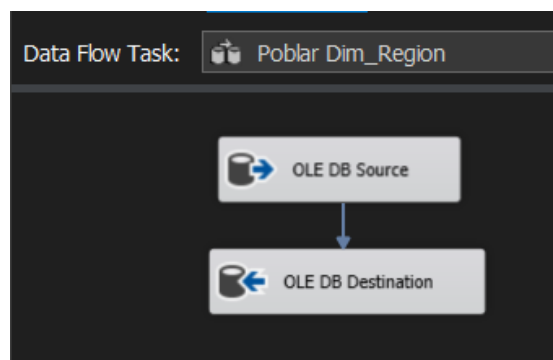
Build Query...

Browse...

Parse Query

Available Input Columns		Available Destination Columns	
Name		Name	
country_region		id_region	
department		country_region	
		department	

Input Column	Destination Column
<ignore>	id_region
country_region	country_region
department	department



Ahora creamos otro flujo para poblar Dim_TipoMovilidad.



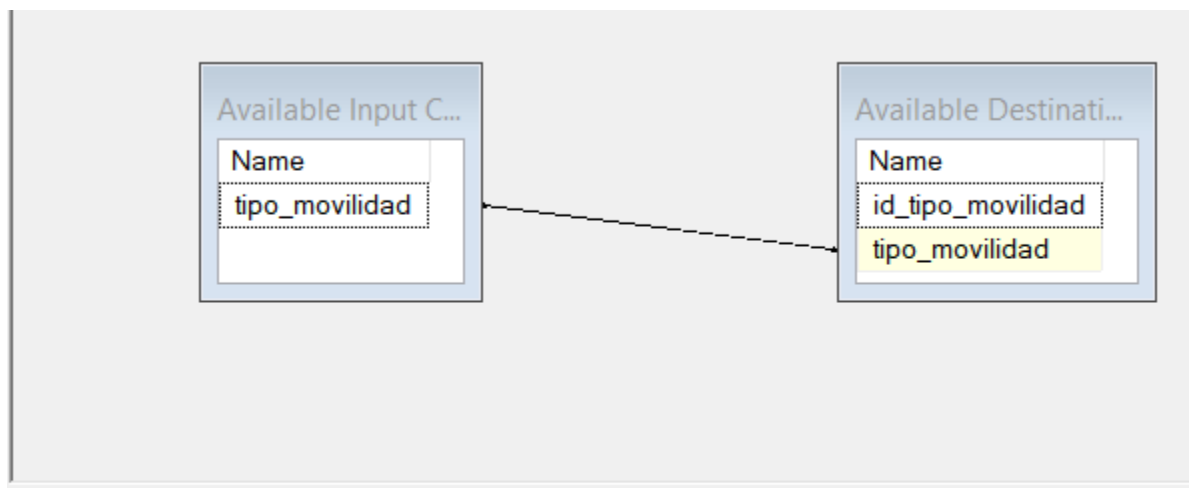
Obtenemos la información con una consulta SQL, y esta la almacenamos en la tabla Dim_TipoMovilidad.

SQL command text:

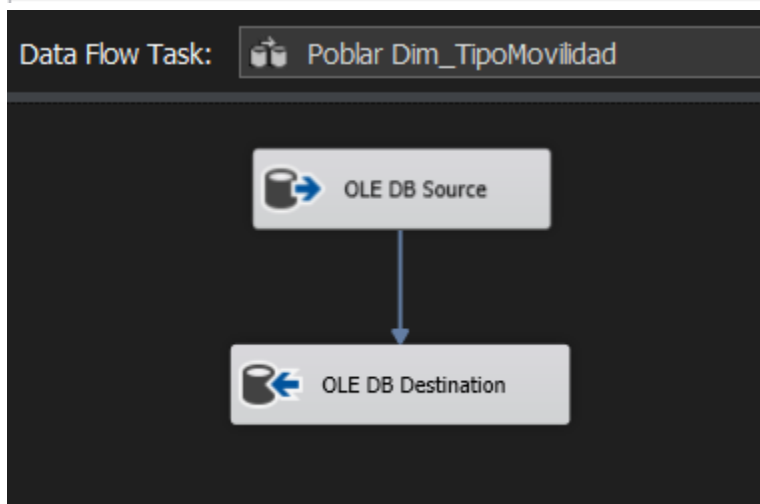
```
SELECT DISTINCT tipo_movilidad
FROM hechos
WHERE tipo_movilidad IS NOT NULL
```

Parameters...

Build Query...



Input Column	Destination Column
<ignore>	id_tipo_movilidad
tipo_movilidad	tipo_movilidad



Finalmente, crearemos el flujo Hechos_Movilidad, el cual creara una tabla que se conectara a las dimensiones previamente creadas. (Adicionalmente se agregara un Execute SQL task para limpiar los datos de las tablas cada vez que ejecutamos el proceso ETL)



En nuestro nuevo flujo llamado Poblar Hechos_Movilidad, creamos un OLE DB Source con la siguiente consulta

SQL command text:

```
SELECT
id_hecho,
mobility_date,
department,
tipo_movilidad,
porcentaje_cambio
FROM hechos
WHERE porcentaje_cambio IS NOT NULL
```

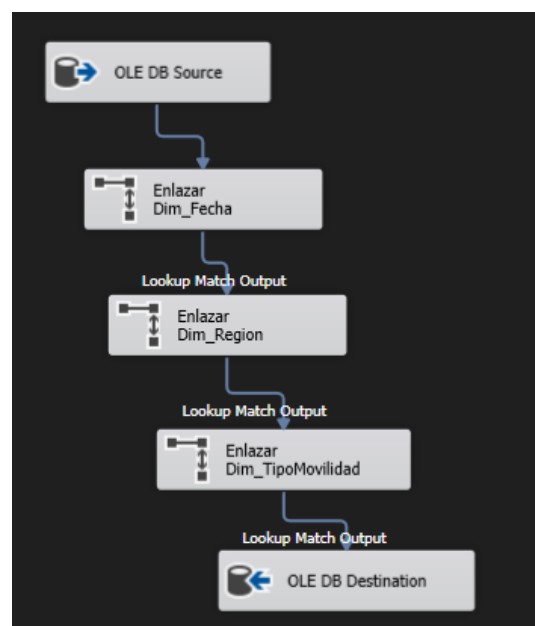
Parameters...

Build Query...

Browse...

Parse Query

Conectamos 3 Lookups y un OLE DB Destination para guardar la tabla, cada uno es para conectar las dimensiones, iremos indagando paso a paso en cada una de las dimensiones.



Conectamos mobility_date de la tabla “hechos” con mobility_date de Dim_Fecha para enlazarlos.

The screenshot shows two panels. The 'Available Input Columns' panel lists 'mobility_date', 'department', 'tipo_movilidad', and 'porcentaje_cambio'. The 'Available Lookup Columns' panel lists 'Name', 'mobility_date', 'year_data', 'month_data', 'day_data', 'month_name', and 'day_month'. A line connects 'mobility_date' in the input panel to 'mobility_date' in the lookup panel.

Lookup Column	Lookup Operation	Output Alias
id_date	<add as new column>	id_date
mobility_date	<add as new column>	mobility_date

Conectamos department de la tabla “hechos” con department de Dim_Region.

The screenshot shows two panels. The 'Available Input Columns' panel lists 'OLE DB Source.mobility_date', 'department', 'tipo_movilidad', 'porcentaje_cambio', 'id_hecho', and 'id_date'. The 'Available Lookup Columns' panel lists 'Name', 'id_region', 'country_region', and 'department'. A line connects 'department' in the input panel to 'department' in the lookup panel.

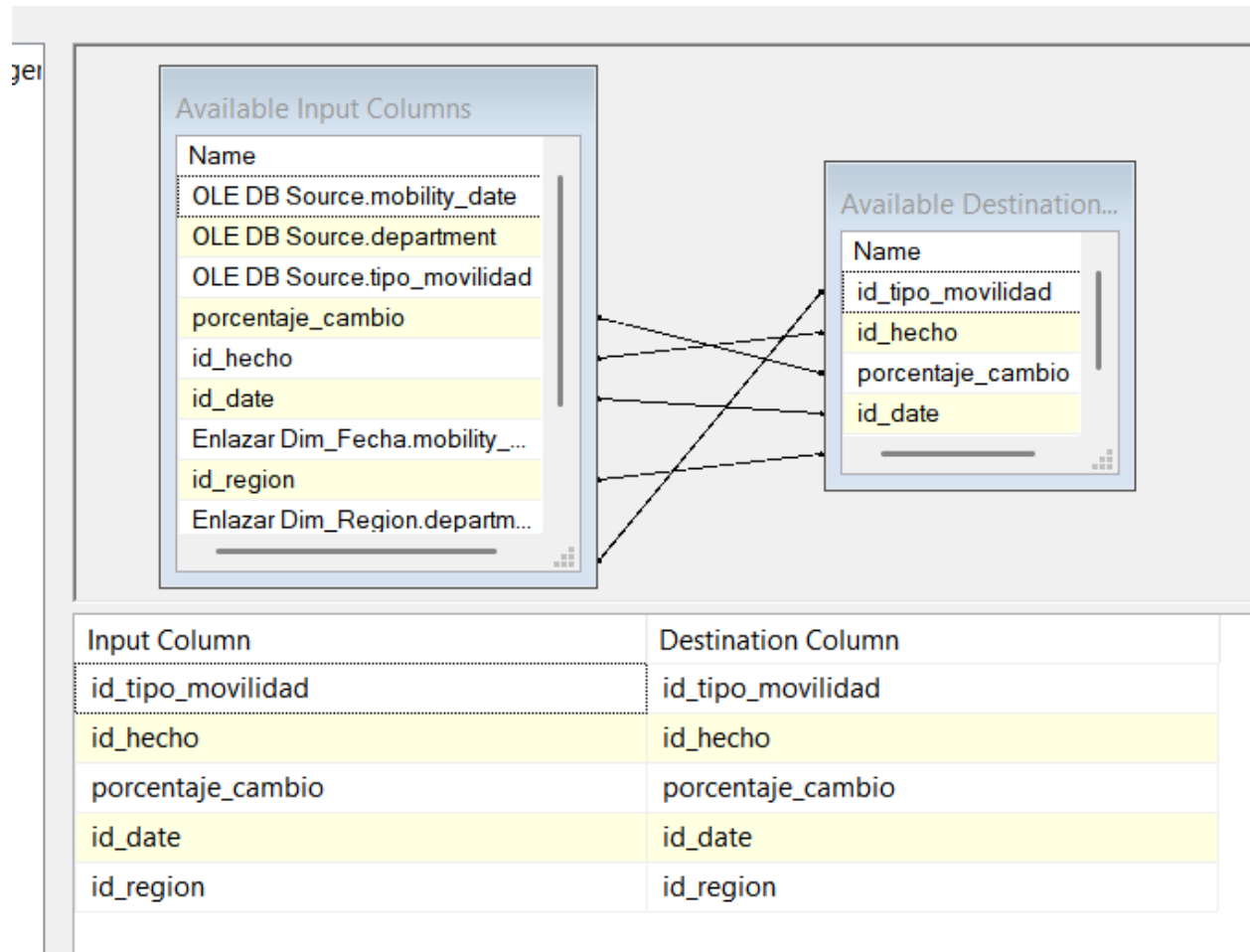
Lookup Column	Lookup Operation	Output Alias
id_region	<add as new column>	id_region
department	<add as new column>	department

Conectamos tipo_movilidad de la tabla “hechos” con tipo_movilidad de Dim_TipoMovilidad.

The screenshot shows two panels. The 'Available Input Columns' panel lists 'OLE DB Source.mobility_date', 'OLE DB Source.department', 'tipo_movilidad', 'porcentaje_cambio', 'id_hecho', 'id_date', 'Enlazar Dim_Fecha.mobility...', and 'id_region'. The 'Available Lookup Columns' panel lists 'Name', 'id_tipo_movilidad', and 'tipo_movilidad'. A line connects 'tipo_movilidad' in the input panel to 'tipo_movilidad' in the lookup panel.

Lookup Column	Lookup Operation	Output Alias
id_tipo_movilidad	<add as new column>	id_tipo_movilidad
tipo_movilidad	<add as new column>	tipo_movilidad

Creamos la tabla “Hechos_Movilidad” y solo conservamos las columnas correspondientes a las llaves foráneas.



Finalmente ejecutamos el proceso ETL

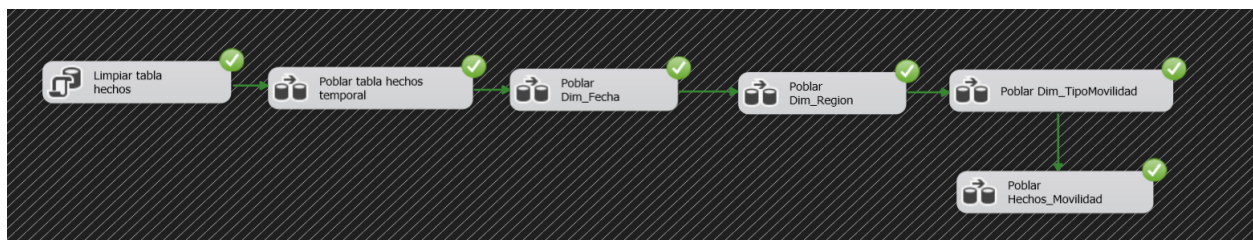
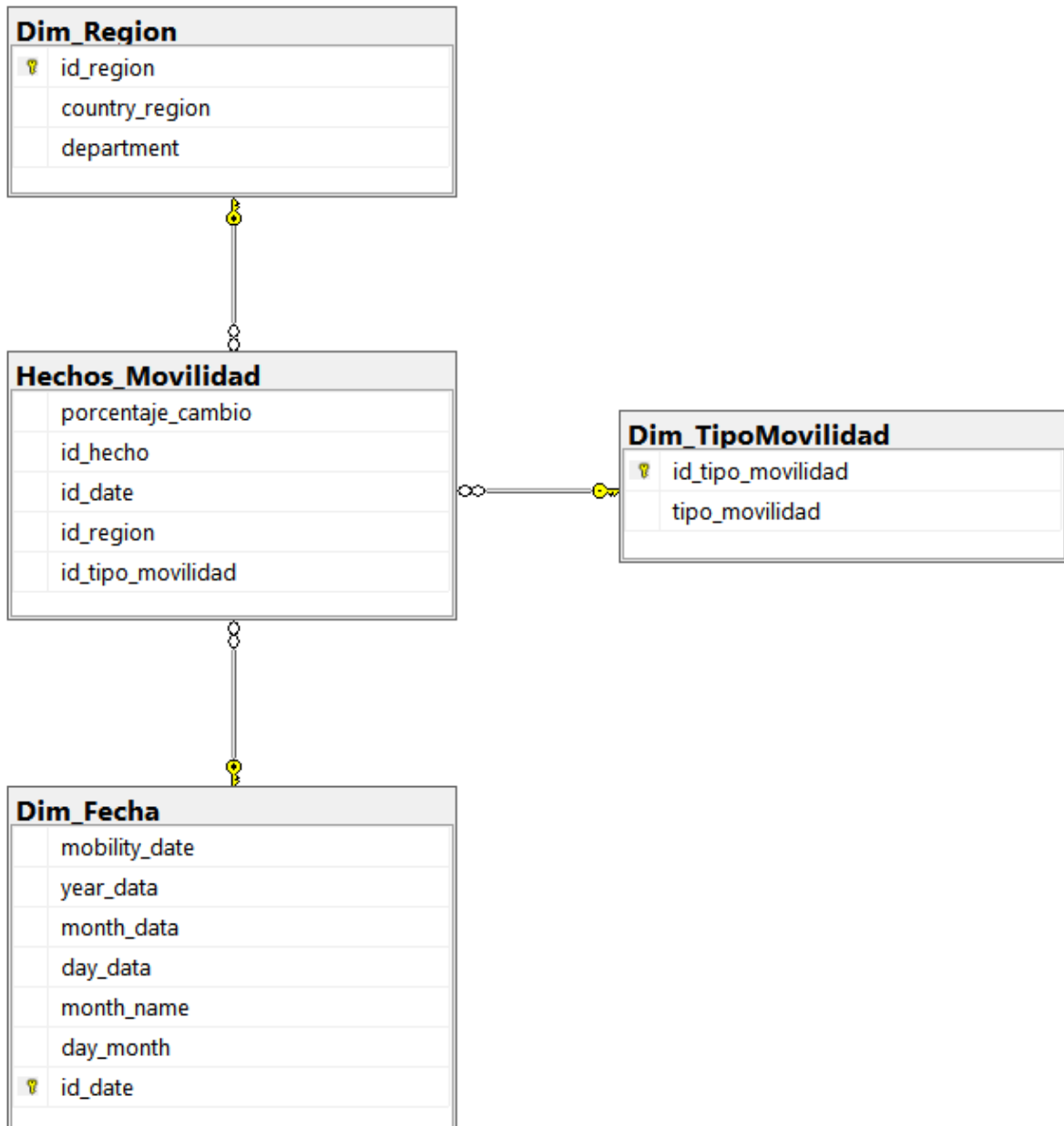
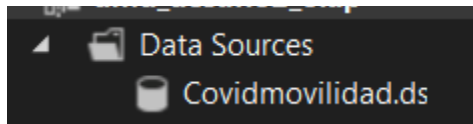


Diagrama de las tablas relacionadas.



Creación de cubo OLAP

Creamos un proyecto de Analysis Services. En primer lugar, crearemos el Data Source Covidmovilidad.ds



Nos conectamos a la base de datos que poblamos con nuestro proceso ETL con el usuario **sa** de SQL Server.

OLE DB Provider:

Microsoft OLE DB Driver for SQL Server ▼ Data Links...

Enter a server or file name

Server or file name: EDUARDO-REPUBLICA

Location:

Log on to the server

☐ Use Windows NT Integrated Security

☒ Use a specific user name and password:

User name: sa

Password:


☐ Blank password

☒ Allow saving password

Initial catalog:

covidmovilidad ▼

Le especificamos que utilice la cuenta de servicio.

 Data Source Designer

General

Impersonation Information

☐ Use a specific Windows user name and password

User name:

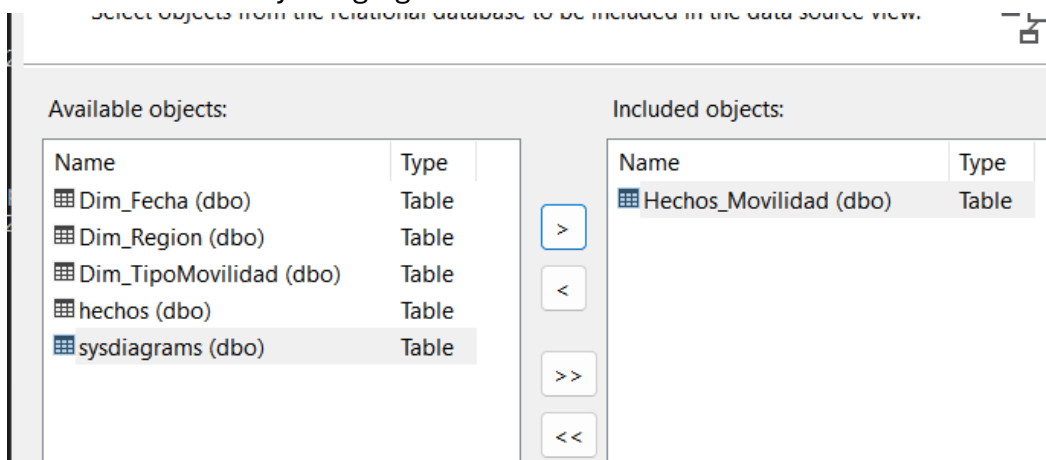
Password:

☒ Use the service account

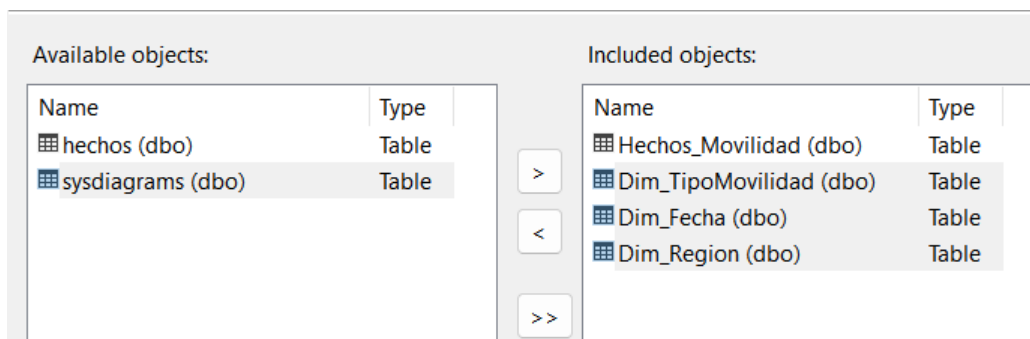
A continuación, crearemos un Data Source View del Data Source que acabamos de crear.

Relational data sources:		Data source properties:	
Covidmovilidad		Property	Value
		Data Sour...	EDUARDO-REPU...
		Initial Cat...	covidmovilidad
		Persist Sec...	True
		Provider	MSOLEDBSQL.1
		User ID	sa

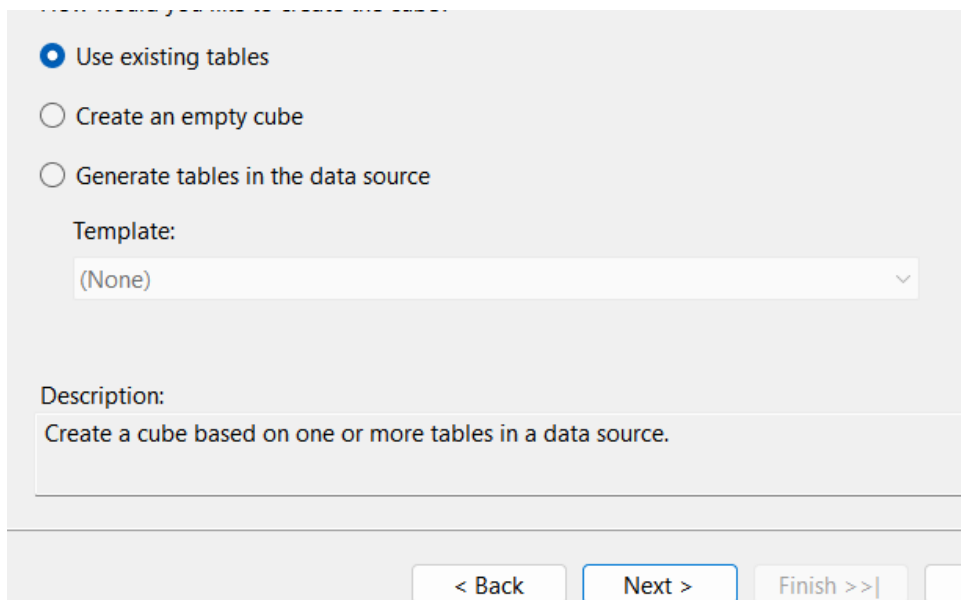
Escogemos que tablas se incluirán, al escoger la tabla de hechos, podemos presionar en Add Related Tables y se agregarán las tablas de dimensiones automáticamente.




Al presionar el botón:



A continuación, crearemos el cubo. Le indicamos que usaremos tablas existentes.



Escogemos la tabla de hechos.





 Cube Wizard

Select Measure Group Tables


Select a data source view or diagram and then select the tables measure groups.

Data source view:
Covidmovilidad

Measure group tables:







- ☒  Hechos_Movilidad
- ☐  Dim_TipoMovilidad
- ☐  Dim_Fecha
- ☐  Dim_Region

Se crean las dimensiones, basadas en las tablas restantes.

 Cube Wizard

Select New Dimensions

Select new dimensions to be created, based on available tables.

- ☒ Dimension
 - ☒  Dim Tipo Movilidad
 - ☒  Dim_TipoMovilidad
 - ☒  Dim Region
 - ☒  Dim_Region
 - ☒  Dim Fecha
 - ☒  Dim_Fecha

Confirmamos y presionamos finalizar.

Cube name:

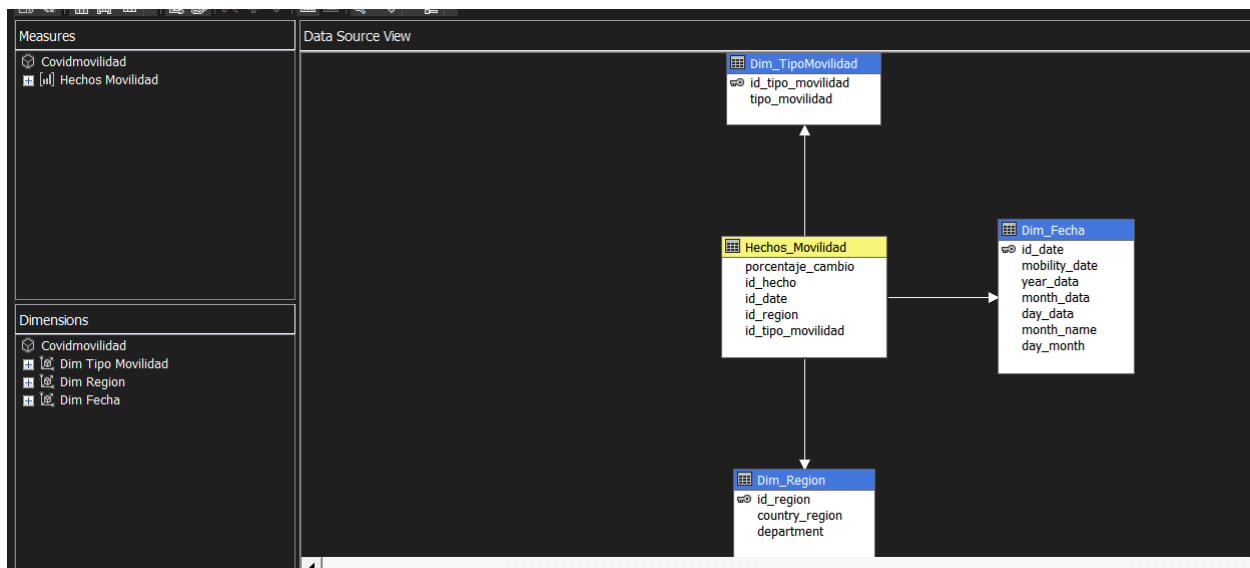
Covidmovilidad

Preview:

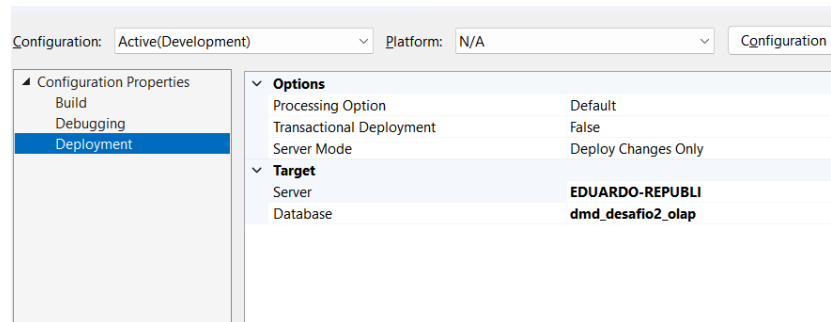
- Measure groups
 - Hechos Movilidad
 - Porcentaje Cambio
 - Id Hecho
 - Hechos Movilidad Count
- Dimensions
 - Dim Tipo Movilidad
 - Dim Region
 - Dim Fecha

< Back Next > Finish Cancel

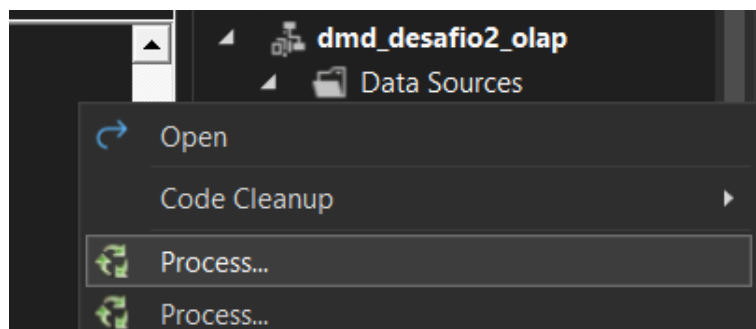
Automáticamente se crea el cubo con sus respectivas dimensiones.



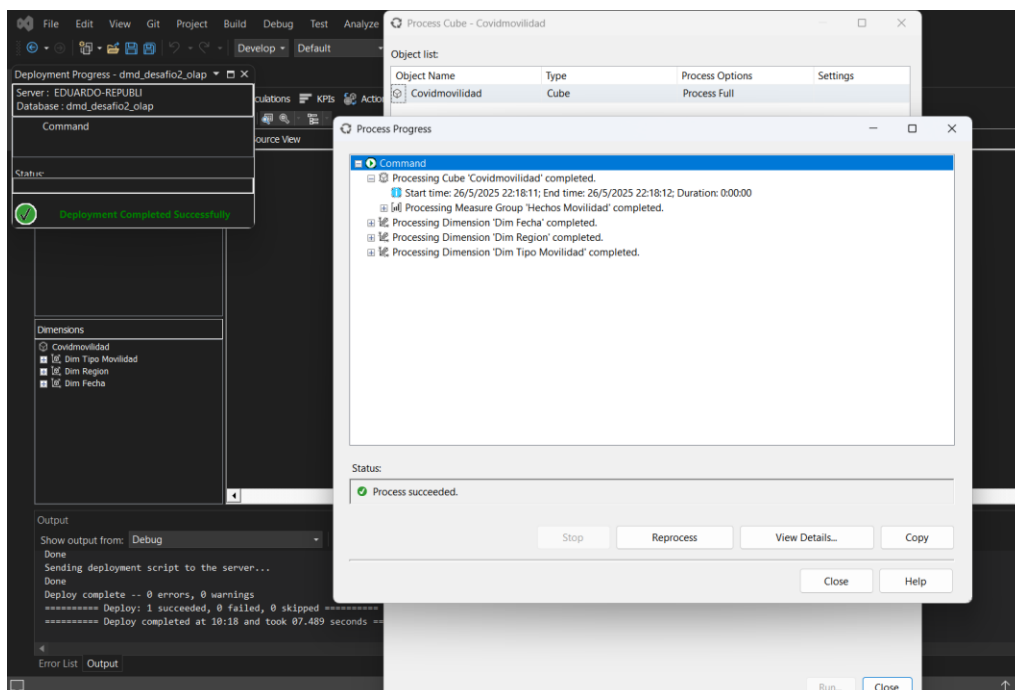
Antes de procesar el cubo, nos vamos a las propiedades de nuestro proyecto, en el apartado de Deployment, en Server reemplazamos “localhost” por el nombre de nuestro servidor de Analysis Services.



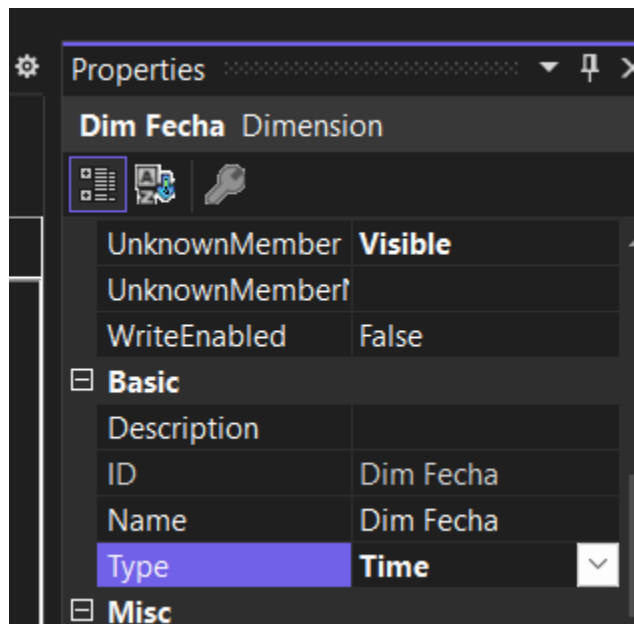
Procedemos a procesar el cubo.



Confirmamos que todo se ejecuto correctamente. Ahora tenemos acceso a mas funcionalidades de nuestro cubo.



Definimos Dim_Fecha como nuestra dimensión de tiempo.



Configuramos el type para cada uno de los campos de Dim_Fecha

