



UNIVERSIDAD DON BOSCO

Desafío #1

Materia: Datawarehouse y Minería de Datos DMD941 G01T

Docente: Ing. Karens Medrano

Integrantes:

Nombres	Apellidos	Carné
Diego Guillermo	Esnard Romero	ER231474
Josue Naum	Dominguez Velasquez	DV241624
Laura Sofía	Pineda Castro	PC230111
Melissa Vanina	López Peña	LP223029
Diego Josué	Padilla Arévalo	PA232942

Fecha de Entrega: 14 de septiembre de 2024

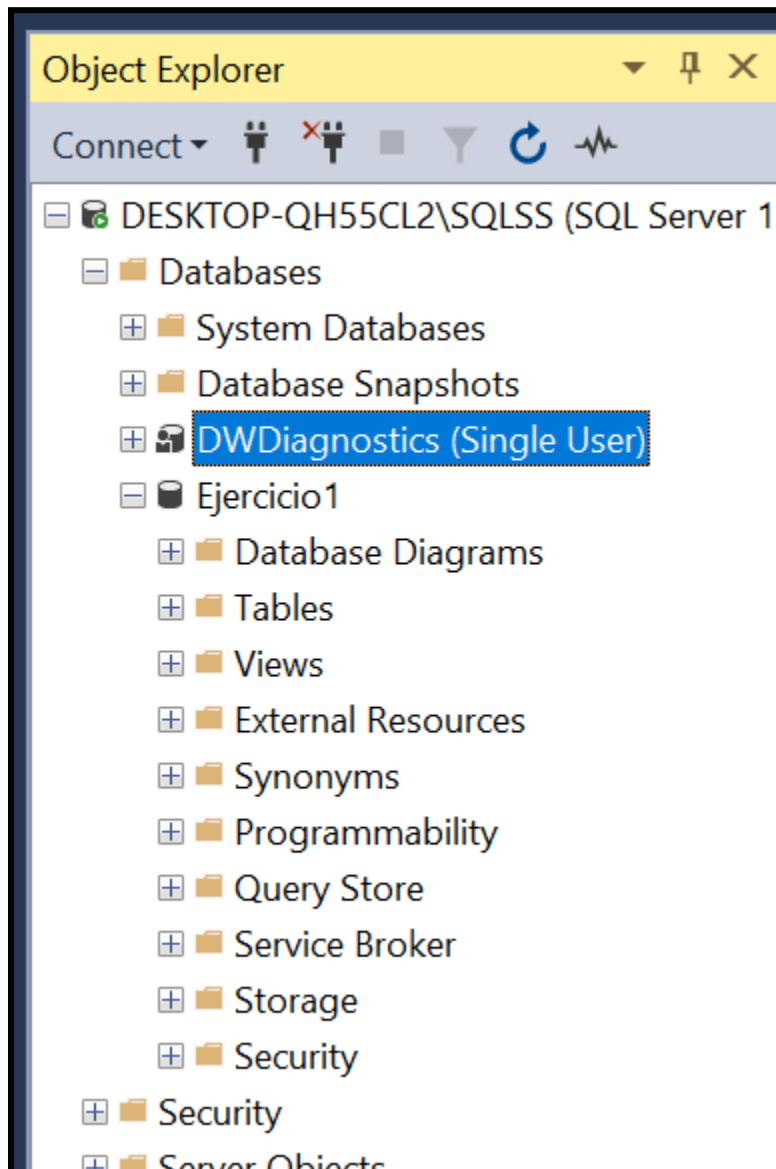
ENLACE A VÍDEO DE LOS EJERCICIOS 1 AL 6:

<https://www.youtube.com/watch?v=sJrELe4L37I>

Ejercicio #1

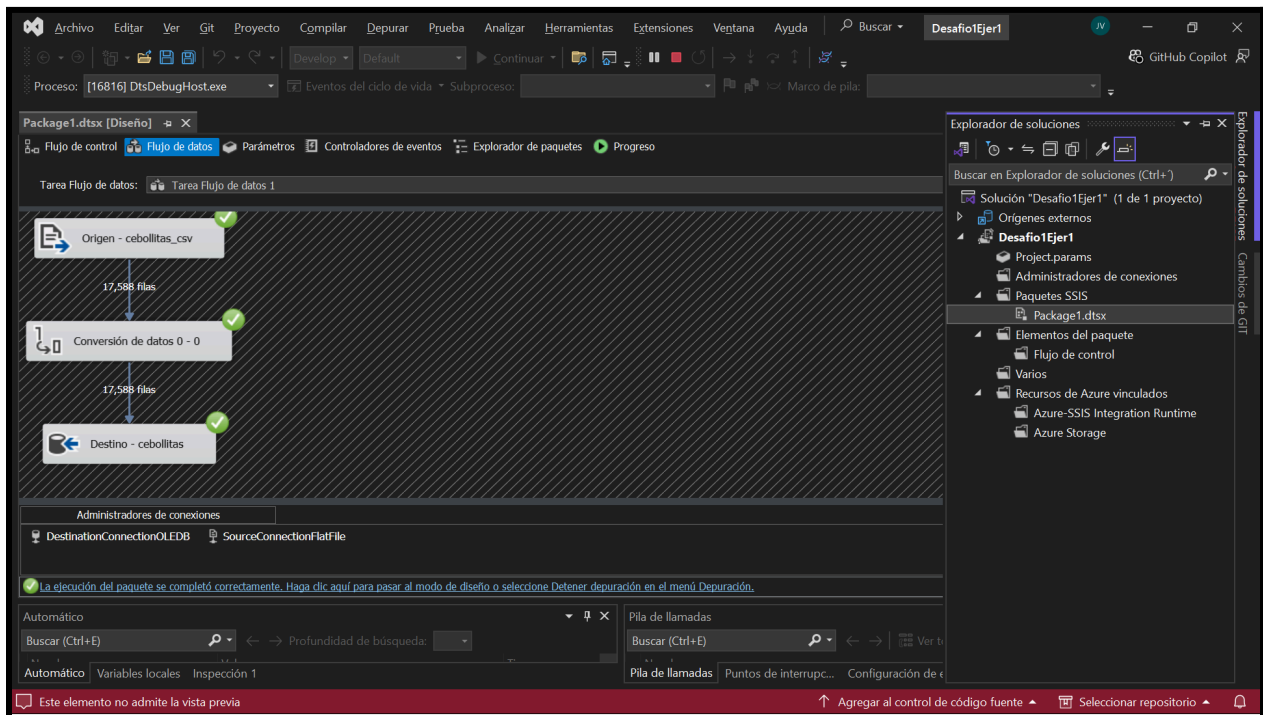
Captura de pantalla de desarrollo de ejercicio 25%

Se crea la base de datos donde se almacenarán las columnas del archivo CSV.



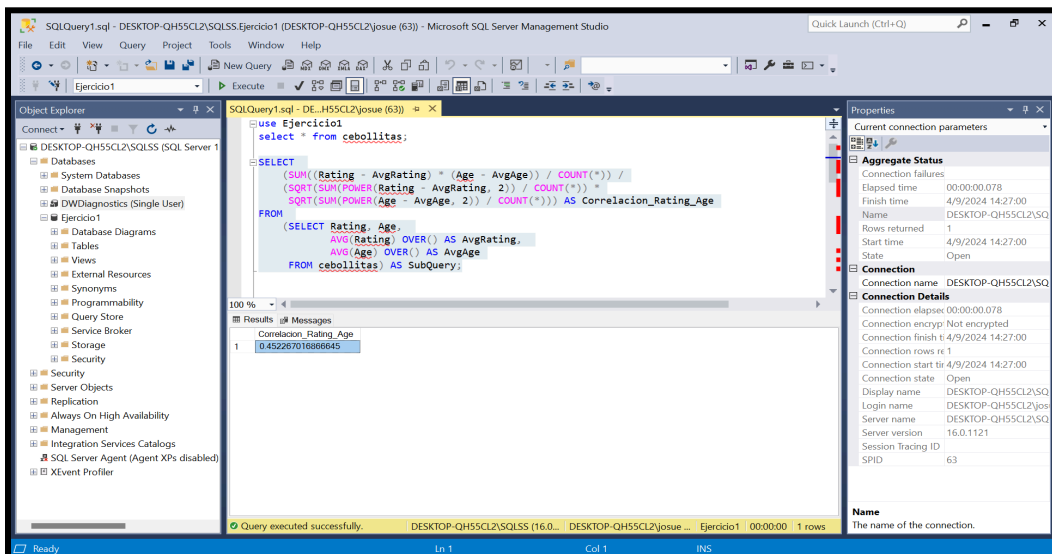
Captura de pantalla de desarrollo de ejercicio 50%

Dentro del Data Flow se coloca un Flat File Source y se configura la ruta del **cebollitas.csv** y se establece una conexión con Connection Manager a la base de datos SQL.



Captura de pantalla de desarrollo de ejercicio 75%

Para calcular la correlación, necesitamos aplicar la fórmula matemática de Pearson en una consulta SQL. Esto nos ayudará a entender cómo el rating está relacionado con la edad de los jugadores. Vamos a escribir la consulta y obtener el resultado.



Captura de pantalla de desarrollo de ejercicio 100%

Generar una tabla que muestre las estadísticas de altura, peso y rating por edad. Esto nos permitirá ver cómo estos parámetros varían con la edad de los jugadores. Redactamos una consulta SQL que agrupe los datos por edad y calcule las estadísticas deseadas para altura, peso y rating. Ejecutamos la consulta para obtener la tabla con la información requerida.

The screenshot displays the Microsoft SQL Server Management Studio interface. The central pane shows a SQL query titled "Convertir y calcular las estadísticas" which uses the `AVG` function to calculate average height, weight, and rating grouped by age. The query is executed against a table named `cebollitas`. The results pane at the bottom shows a table with 10 rows of data, including age, average height, average weight, and average rating. The status bar at the bottom indicates the query was executed successfully, returning 29 rows.

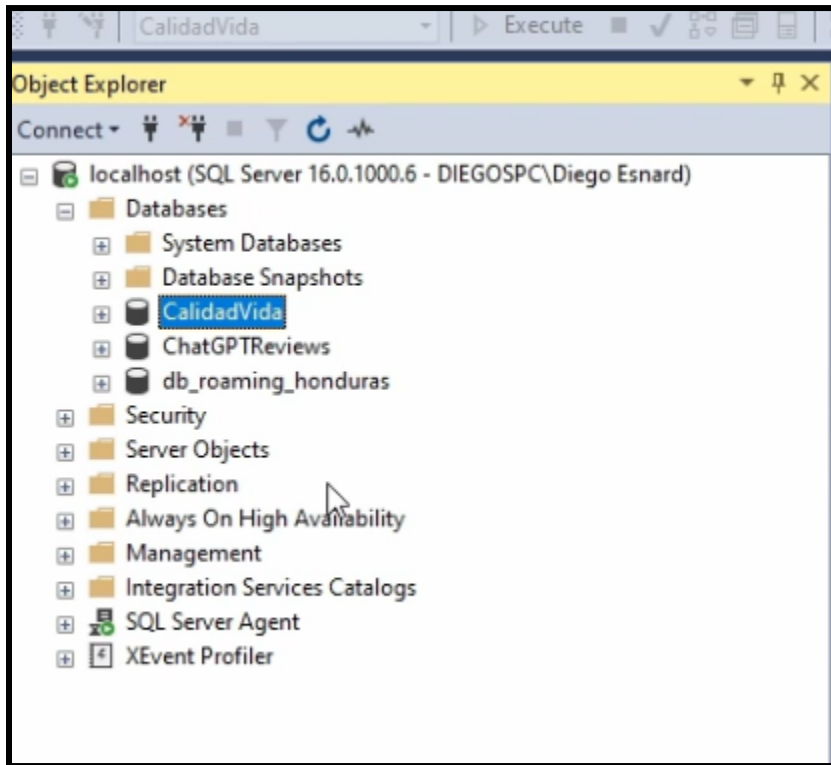
```
-- Convertir y calcular las estadísticas
SELECT
    Age,
    AVG(CAST(REPLACE(REPLACE(Height, ' cm', ''), ', ', '')) AS FLOAT) AS Avg_Height,
    AVG(CAST(REPLACE(REPLACE(Weight, ' kg', ''), ', ', '')) AS FLOAT) AS Avg_Weight,
    AVG(Rating) AS Avg_Rating
INTO
    PlayerStatsByAge
FROM
    cebollitas
GROUP BY
    Age
ORDER BY
    Age;
select * from PlayerStatsByAge
```

	Age	Avg_Height	Avg_Weight	Avg_Rating
1	23	180.915191740413	74.7927728613569	65
2	29	181.126811594203	76.1757246376812	69
3	32	181.562610229277	76.8465608465608	69
4	26	181.483682008368	76.2435146443515	68
5	35	182.063559322034	77.9872881355932	69
6	43	182	82.6666666666667	60
7	27	180.930335097002	75.7451499118166	68
8	21	180.416387959866	73.5008361204013	62
9	38	181.54	77.4	66
10	44	182.222222222222	82.2222222222222	67

Ejercicio #2

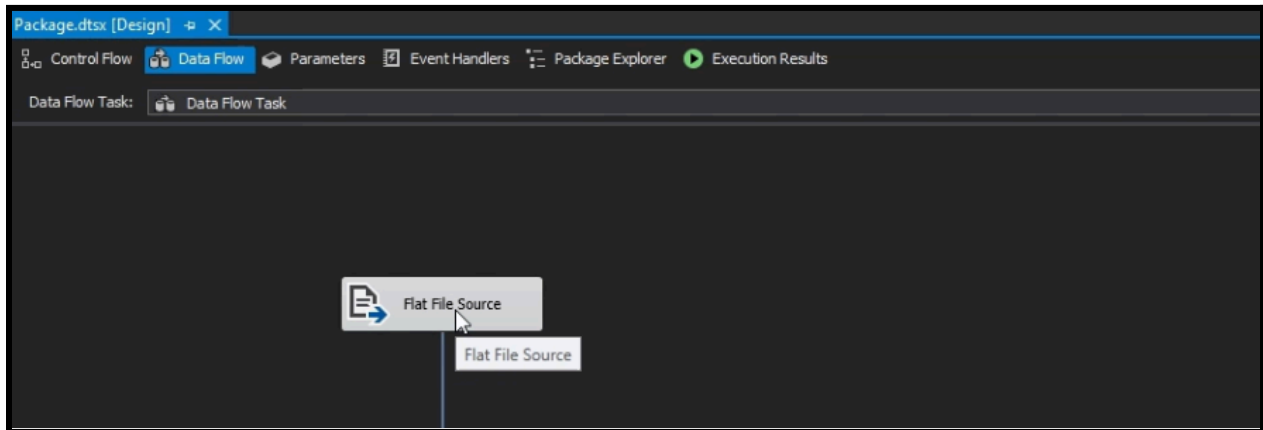
Captura de pantalla de desarrollo de ejercicio 25%

Creación de base de datos para almacenar los datos del archivo CSV con sus respectivas columnas.



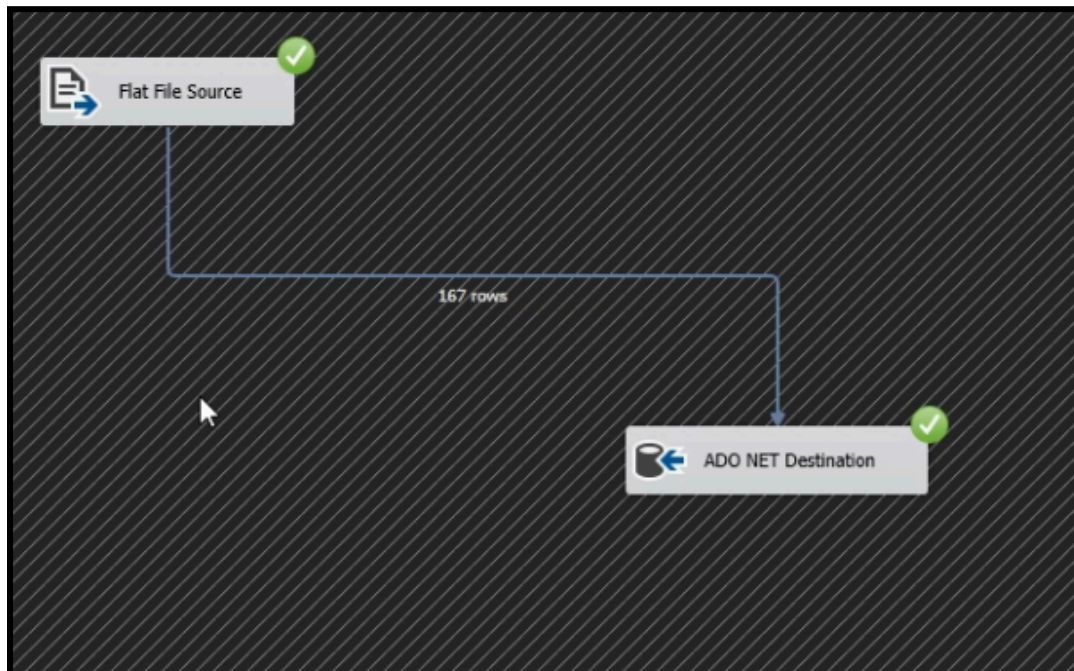
Captura de pantalla de desarrollo de ejercicio 50%

Creación de un Data Flow Task y dentro de este mismo, arrastrar un Flat File Source con la ubicación del archivo CSV para exportar todos los datos.



Captura de pantalla de desarrollo de ejercicio 75%

Se coloca un ADO NET Destination donde se configura la ruta de la base de datos SQL y se une al Flat File Source del archivo CSV. Se hace un mapeo de las columnas del archivo CSV y las columnas creadas de la base de datos SQL. Se ejecuta para corroborar que toda la información se pasa exitosamente desde el archivo CSV a la base de datos SQL.



Captura de pantalla de desarrollo de ejercicio 100%

Se crea una nueva columna llamada Continente y se agrega a cada país al tipo de continente que pertenece, luego se realizan las consultas correspondientes para mostrar en pantalla lo solicitado.

The screenshot shows a SQL query editor with four queries. The first query is selected, and its results are displayed in a table below. The table has five columns: Pais, Salud, Educacion, CalidadEconomica, and CondicionesVida. The results are sorted by Pais in descending order.

```
--Consulta para obtener los países con las mejores condiciones de vida--
SELECT Pais, PuntuacionPromedio
FROM DatosCalidadVida
ORDER BY PuntuacionPromedio DESC;

--Comparar la seguridad y libertad personal por país--
SELECT Pais, IndiceSeguridadPersonal, IndiceLibertad
FROM DatosCalidadVida
ORDER BY IndiceSeguridadPersonal DESC, IndiceLibertad DESC;

--Calcular el promedio de las puntuaciones por continente--
SELECT Continente,
       ROUND(AVG(PuntuacionPromedio), 2) AS PromedioCalidadVida,
       ROUND(AVG(IndiceSeguridadPersonal), 2) AS PromedioSeguridadPersonal,
       ROUND(AVG(IndiceLibertad), 2) AS PromedioLibertad
FROM DatosCalidadVida
WHERE Continente IS NOT NULL
GROUP BY Continente;

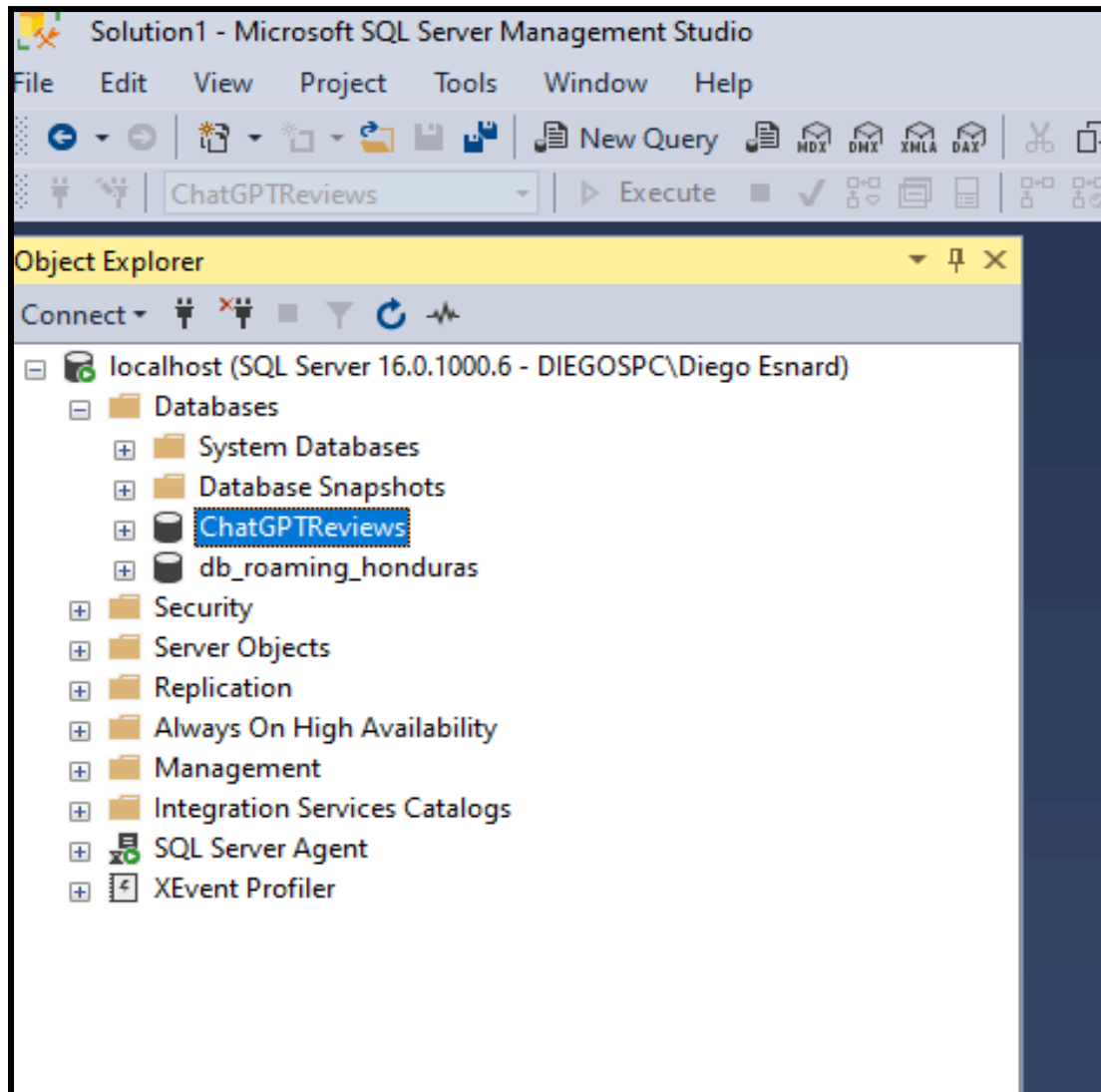
--Comparar otros indicadores de calidad de vida--
SELECT Pais, Salud, Educacion, CalidadEconomica, CondicionesVida
FROM DatosCalidadVida
ORDER BY Salud DESC, Educacion DESC;
```

	Pais	Salud	Educacion	CalidadEconomica	CondicionesVida
1	Singapore	86.89	91.44	80.1	93.35
2	Japan	86.5	84.93	66.35	92.86
3	SouthKorea	84.8	87.76	74.59	91.47
4	Taiwan	83.37	84.61	73.86	90.22
5	China	83.11	70.42	68.46	85.05
6	Israel	83.1	81.63	70.96	93.5
7	Norway	82.98	85.68	77.25	94.7
8	Iceland	82.72	85.19	69.92	93.82
9	Sweden	82.28	85.92	76.18	95.33
10	Switzerland	82.11	87.72	79.71	94.66
11	Netherlands	82.05	86.43	74.34	95.86
12	Luxembourg	81.59	78.79	76.93	94.56

Ejercicio #3

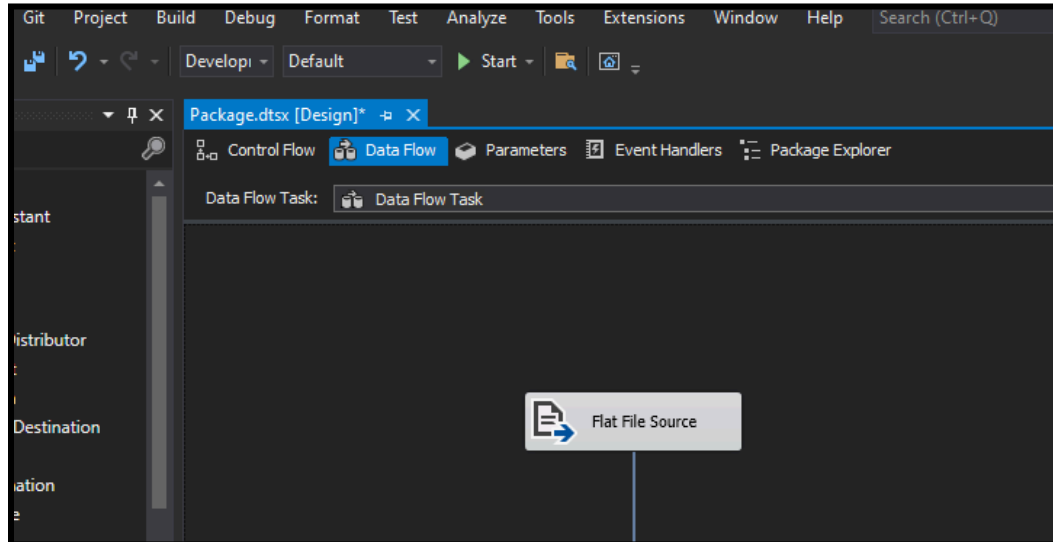
Captura de pantalla de desarrollo de ejercicio 25%

Se crea la base de datos donde se almacenarán las columnas del archivo CSV.



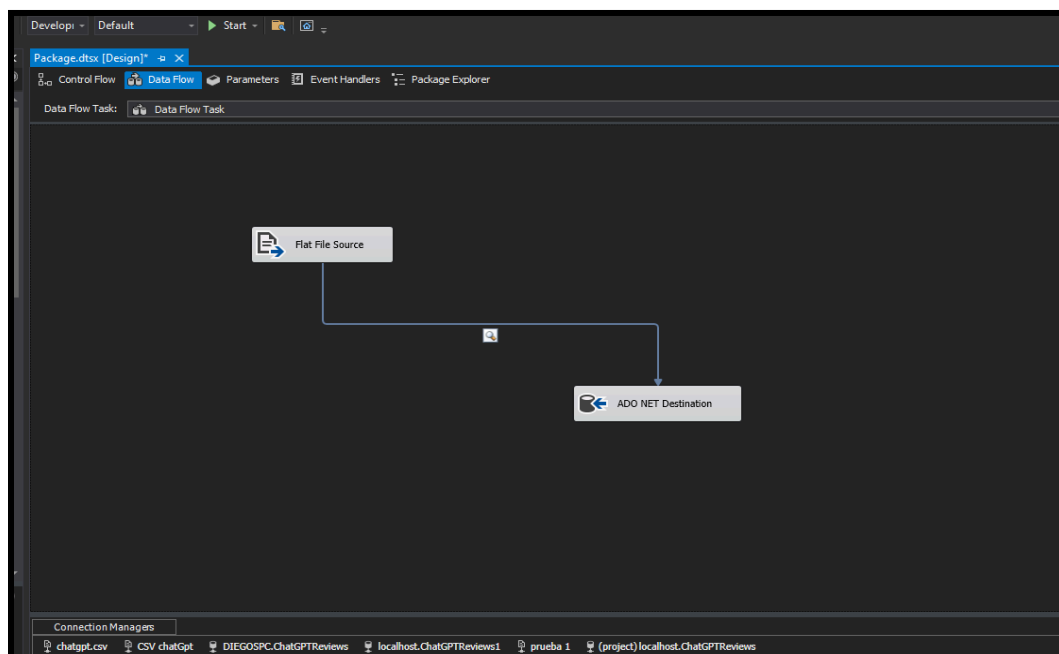
Captura de pantalla de desarrollo de ejercicio 50%

Dentro del Data Flow se coloca un Flat File Source y se configura la ruta del **ChatGpt.csv** y se establece una conexión con Connection Manager a la base de datos SQL.



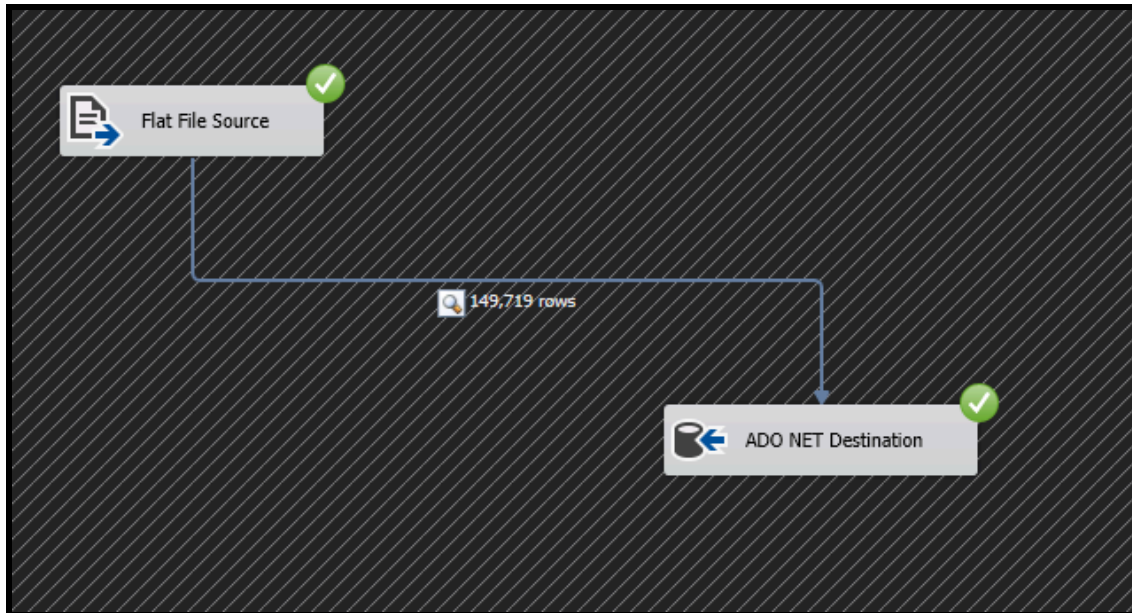
Captura de pantalla de desarrollo de ejercicio 75%

Se coloca un ADO NET Destination donde se configura la ruta de la base de datos SQL y se une al Flat File Source del archivo CSV.



Captura de pantalla de desarrollo de ejercicio 100%

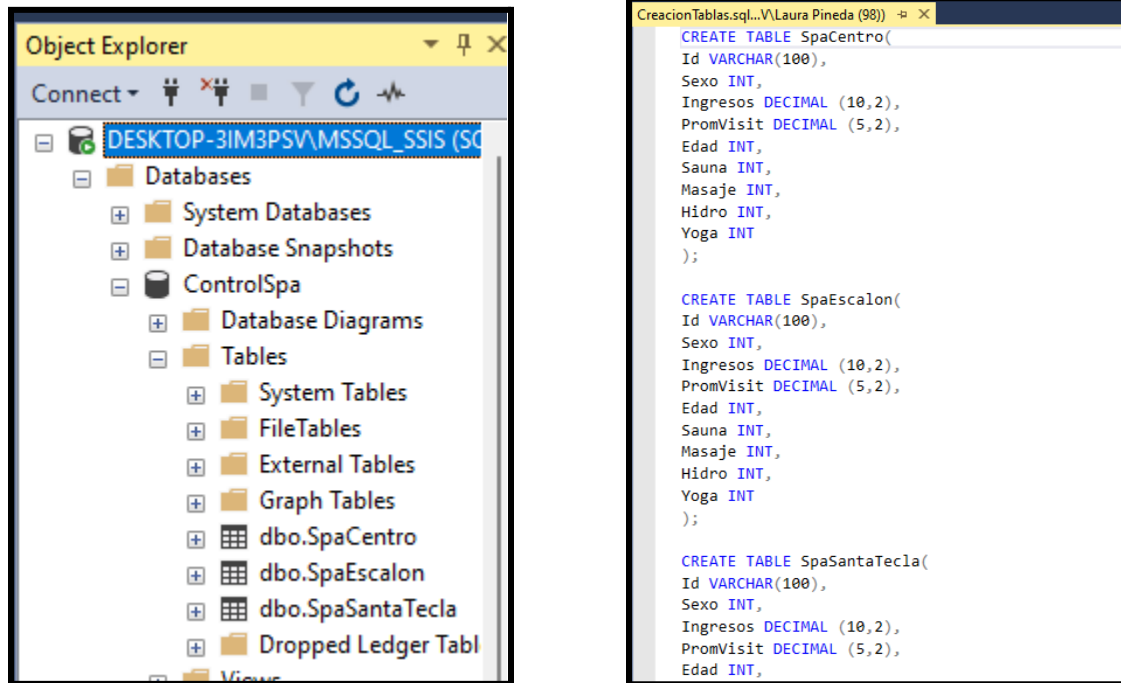
Se ejecuta para confirmar que los datos del archivo CSV se están trasladando correctamente a la base de datos y se corrobora en la base de datos SQL que las columnas del archivo CSV están en la nueva base de datos creada al inicio.



Ejercicio #4

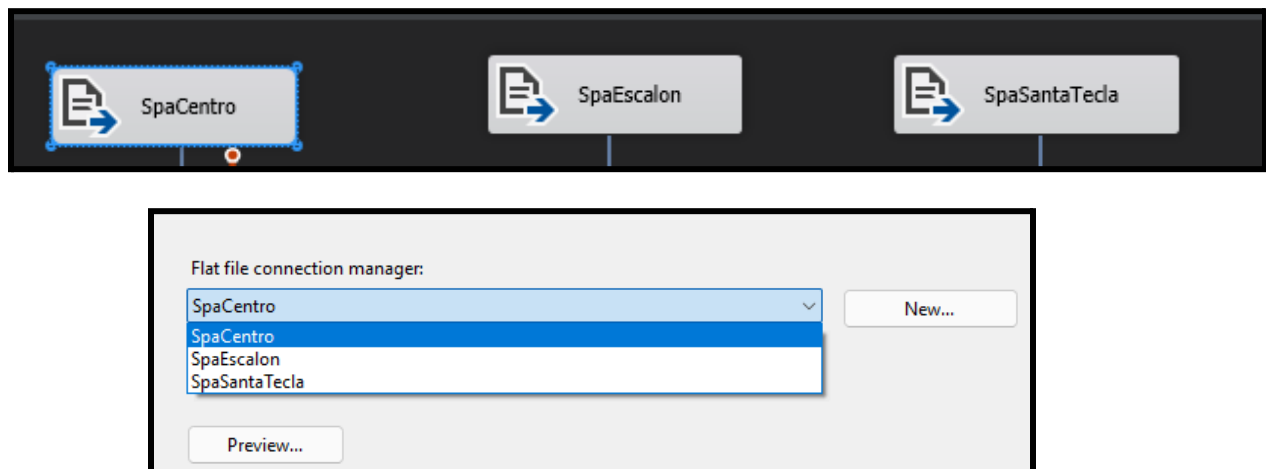
Captura de pantalla de desarrollo de ejercicio 25%

Se crea la base de datos y las tablas correspondientes donde se almacenará cada archivo CSV.



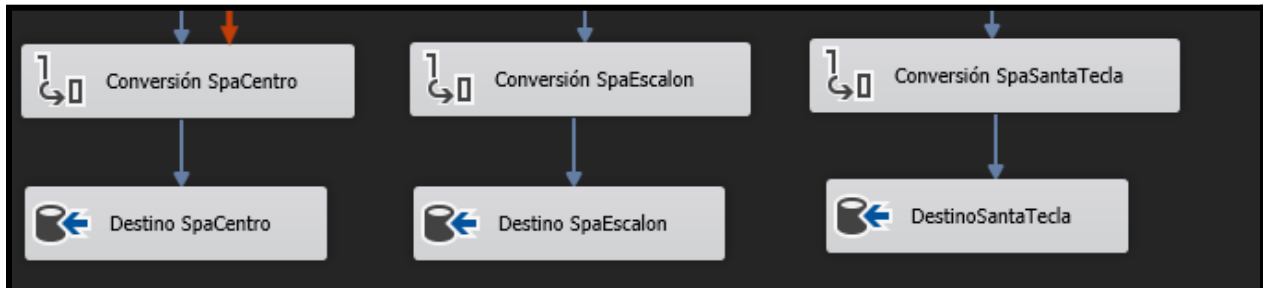
Captura de pantalla de desarrollo de ejercicio 50%

Agregamos un 'Data Flow' y en él se agregan 3 'Flat File Source' En donde agregamos los 3 archivos CSV de los cuales se extrae la información.

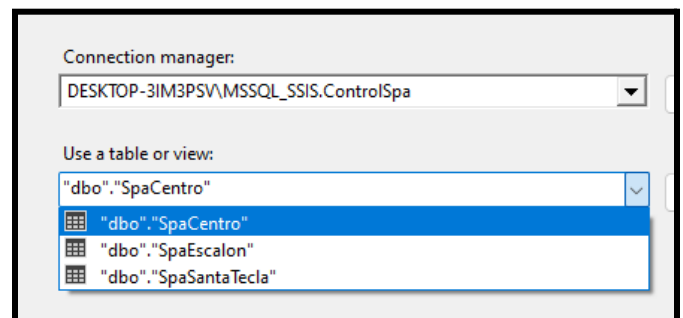


Captura de pantalla de desarrollo de ejercicio 75%

Se coloca un Conversor de datos para convertir los datos del CSV a datos 'UTF-8' y luego agregamos un 'ADO NET Destination' donde se configura la base de datos SQL en donde se guardarán los datos y sus respectivas tablas.

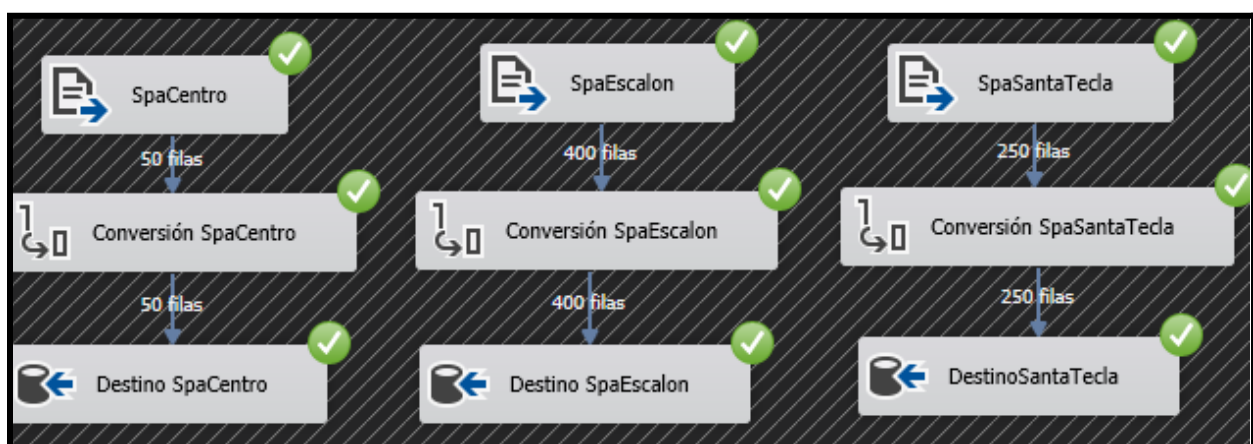


Input Column	Output Al...	Data Type	Length	P. S...	Code Page
id	Copy of id	cadena [DT_STR]	50		65001 (UTF-8)
Sexo	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
ingresos	Copy of i...	cadena [DT_STR]	50		65001 (UTF-8)
PromVisit	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
Edad	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
Sauna	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
Masaje	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
Hidro	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)
Yoga	Copy of ...	cadena [DT_STR]	50		65001 (UTF-8)



Captura de pantalla de desarrollo de ejercicio 100%

Se ejecuta para confirmar que los datos del archivo CSV se están trasladando correctamente a la base de datos y para corroborar procedemos a implementar las segmentaciones según deseemos



```

SELECT
    Id,
    CASE
        WHEN Edad < 25 THEN 'Jóvenes'
        WHEN Edad BETWEEN 25 AND 45 THEN 'Adultos'
        ELSE 'Adultos Mayores'
    END AS GrupoEdad
FROM (
    SELECT Id, Edad FROM SpaCentro
    UNION ALL
    SELECT Id, Edad FROM SpaEscalon
    UNION ALL
    SELECT Id, Edad FROM SpaSantaTecla
) AS Clientes;

```

	Id	GrupoEdad
1	Tomkin Stickles	Jóvenes
2	Tyson Stovine	Adultos
3	Miller Camachen	Jóvenes
4	Damell Dine-Hart	Adultos Mayores
5	Wyatt Keyte	Adultos
6	Trip Vost	Adultos Mayores
7	Ammamaria D'Erico	Adultos
8	Jessica Kuhn	Adultos Mayores
9	Hamnet Lindenberg	Adultos Mayores
10	Marga Kiley	Adultos
11	Pebrook Praton	Adultos
12	Koenraad Marchiso	Adultos

```

SELECT
    Id,
    CASE
        WHEN Ingresos > 2000 AND PromVisit > 5 THEN 'Premium'
        WHEN Ingresos BETWEEN 1000 AND 2000 AND PromVisit BETWEEN
        3 AND 5 THEN 'Regular'
        ELSE 'Ocasional'
    END AS SegmentoCliente,
    CASE
        WHEN Edad < 25 THEN 'Jóvenes'
        WHEN Edad BETWEEN 25 AND 45 THEN 'Adultos'
        ELSE 'Adultos Mayores'
    END AS GrupoEdad
FROM (
    SELECT Id, Ingresos, PromVisit, Edad FROM SpaCentro
    UNION ALL
    SELECT Id, Ingresos, PromVisit, Edad FROM SpaEscalon
    UNION ALL
    SELECT Id, Ingresos, PromVisit, Edad FROM SpaSantaTecla
) AS Clientes;

```

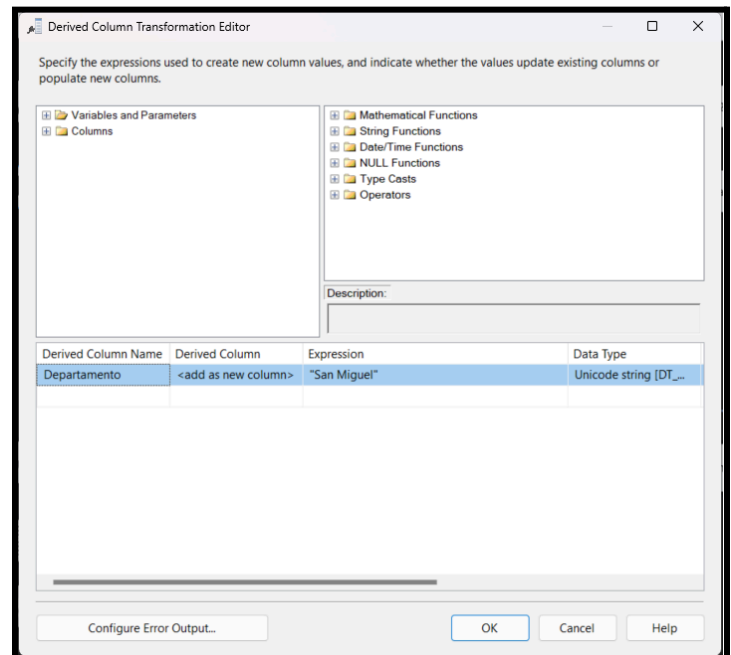
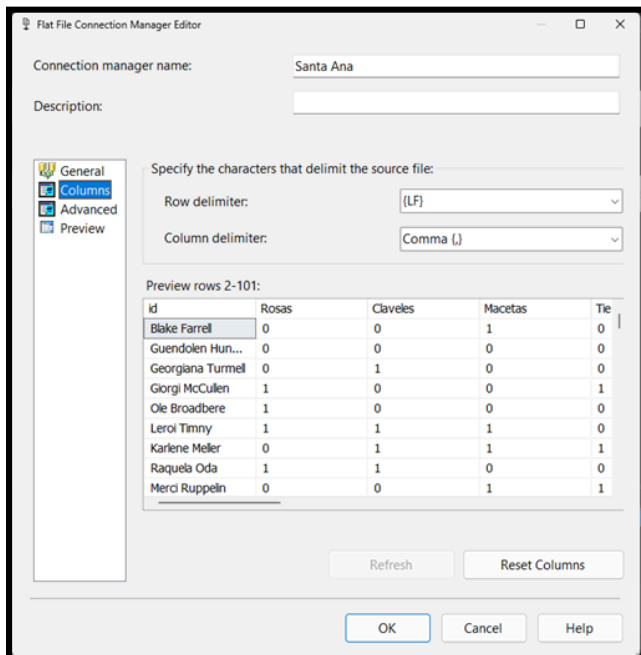
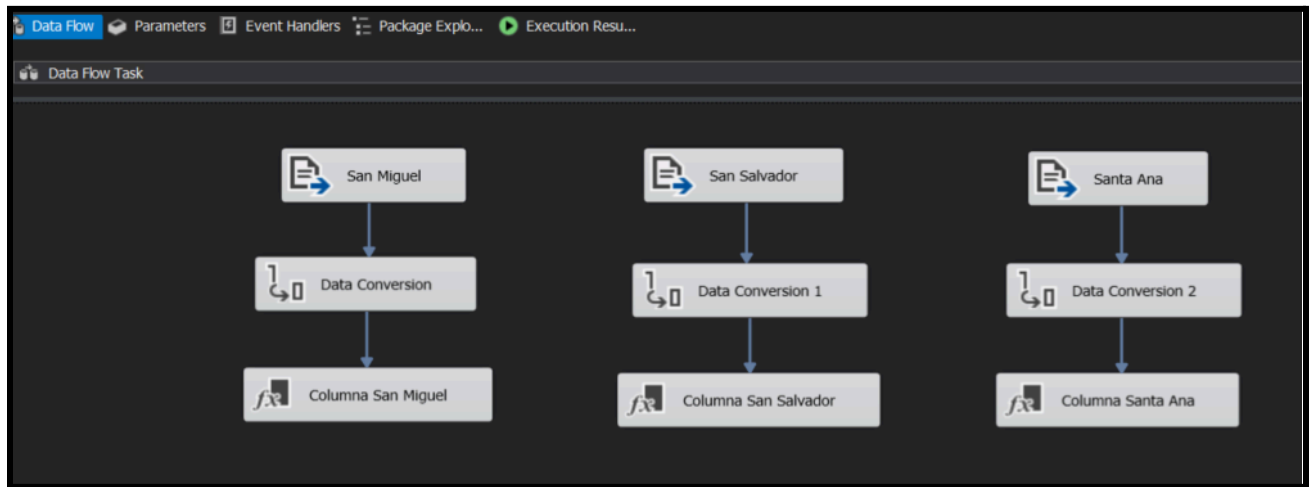
110 %

	Id	SegmentoCliente	GrupoEdad
1	Tomkin Stickles	Ocasional	Jóvenes
2	Tyson Stovine	Premium	Adultos
3	Miller Camachen	Ocasional	Jóvenes
4	Damell Dine-Hart	Regular	Adultos Mayores
5	Wyatt Keyte	Ocasional	Adultos
6	Trip Vost	Ocasional	Adultos Mayores
7	Ammamaria D'Erico	Ocasional	Adultos
8	Jessica Kuhn	Ocasional	Adultos Mayores
9	Hamnet Lindenberg	Ocasional	Adultos Mayores
10	Marga Kiley	Ocasional	Adultos
11	Pebrook Praton	Ocasional	Adultos
12	Koenraad Marchiso	Ocasional	Adultos

Ejercicio #5

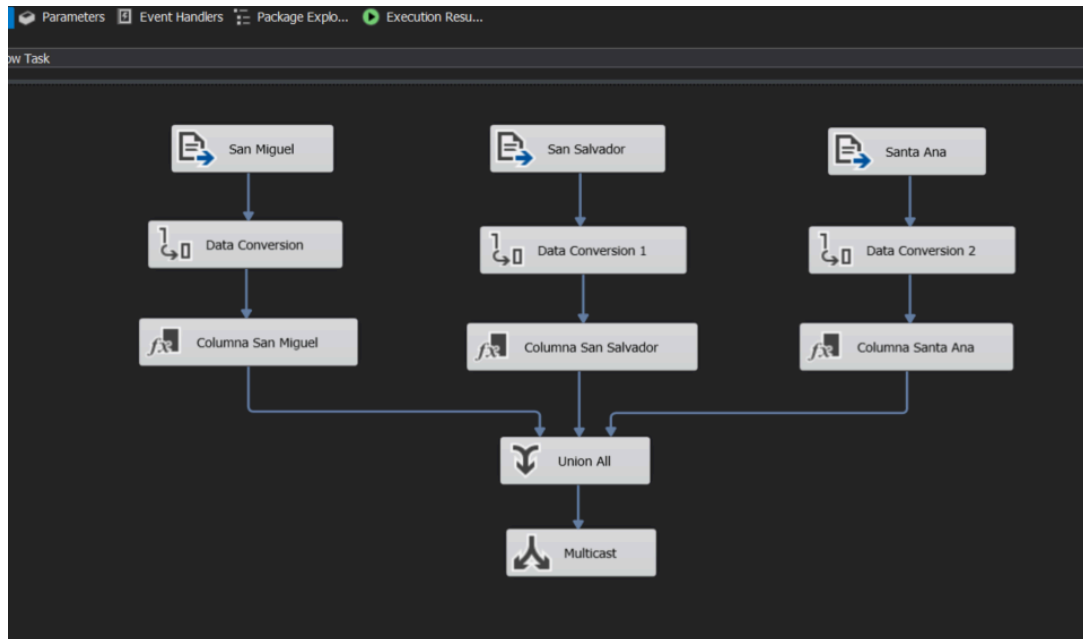
Captura de pantalla de desarrollo de ejercicio 25%

Luego de agregar el Control Flow, se cargan los archivos CSV de cada departamento como Flat File Source en el Data Flow. Se realiza la conversión de datos del ID para que todos sean Unicode string [DT_WSTR] y se crea una nueva columna para agregar el departamento al que pertenece venta realizada.



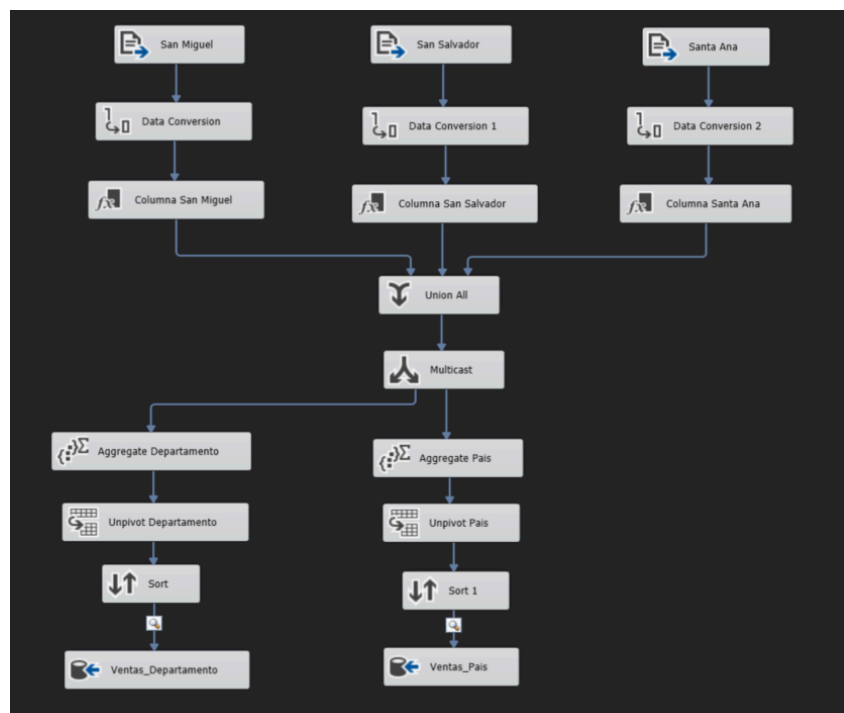
Captura de pantalla de desarrollo de ejercicio 50%

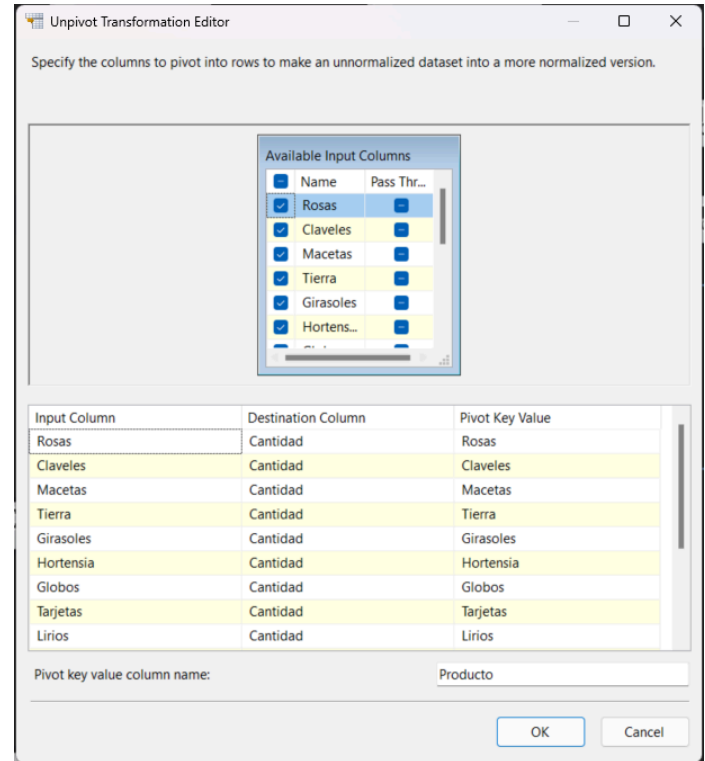
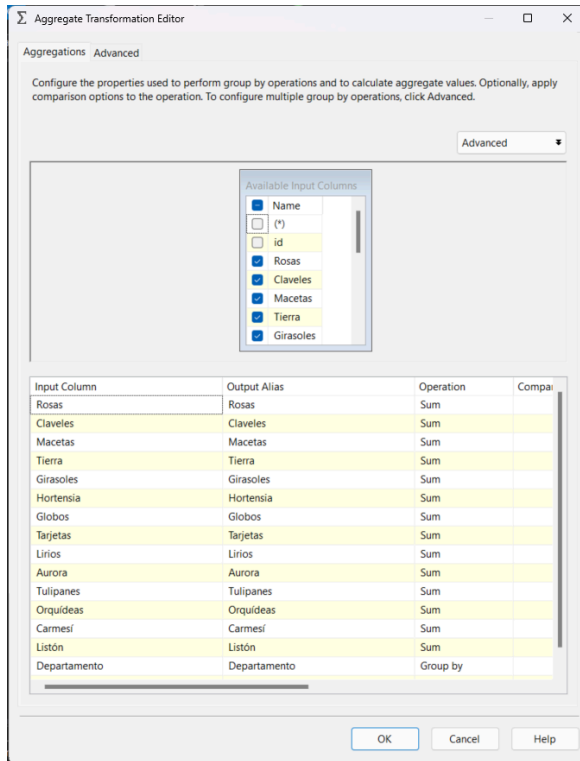
Se implementa la transformación Union All para combinar los datos de los tres departamentos y luego se hace uso de un Multicast para procesar los datos según departamento y país.



Captura de pantalla de desarrollo de ejercicio 75%

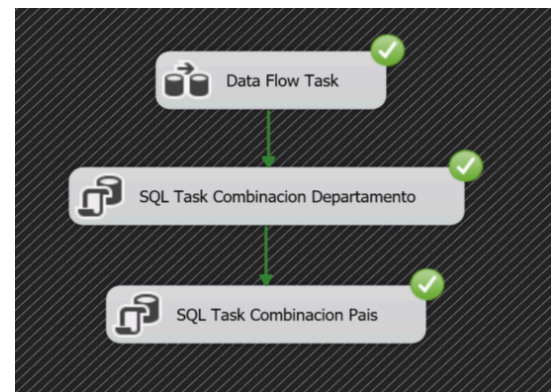
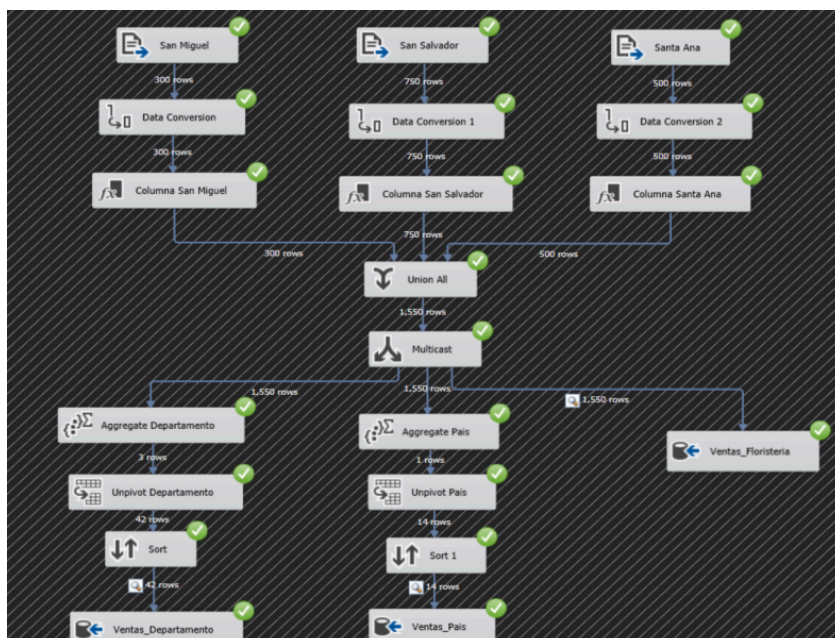
Se implementa un Aggregate para sumar las ventas por cada producto por departamento y país. Luego se usa la transformación Unpivot para transformar los datos de columnas (Rosas, Tulipanes, etc) en filas y facilitar el análisis según la cantidad vendida. Se usa Sort para ordenar los productos según los más vendidos por departamento y país.





Captura de pantalla de desarrollo de ejercicio 100%

Se agrega un Destino OLE DB para exportar los datos unidos a SQL. Luego, en el Control Flow se agrega un SQL Task que contiene un query para determinar las mejores combinaciones de productos por departamento y otro para las mejores combinaciones por país. Finalmente se ejecuta y se analizan todos los datos obtenidos.



Productos que sobresalen por Departamento y País

Producto	Cantidad	Departamento
Aurora	160	San Miguel
Lirios	160	San Miguel
Carmesí	158	San Miguel
Orquídeas	158	San Miguel
Rosas	157	San Miguel
Hortensia	157	San Miguel
Globos	151	San Miguel
Girasoles	150	San Miguel
Tulpanes	149	San Miguel
Listón	149	San Miguel
Tarjetas	143	San Miguel
Macetas	141	San Miguel
Tierra	141	San Miguel
Claveles	137	San Miguel
Listón	690	San Salvador
Rosas	612	San Salvador
Globos	587	San Salvador
Macetas	392	San Salvador
Aurora	384	San Salvador
Tarjetas	384	San Salvador
Orquídeas	380	San Salvador
Hortensia	374	San Salvador
Girasoles	371	San Salvador
Tierra	368	San Salvador
Lirios	365	San Salvador
Tulpanes	357	San Salvador
Carmesí	353	San Salvador
Claveles	350	San Salvador
Lirios	270	Santa Ana
Girasoles	266	Santa Ana
Aurora	260	Santa Ana
Orquídeas	259	Santa Ana
Tarjetas	252	Santa Ana
Tulpanes	247	Santa Ana
Claveles	246	Santa Ana
Macetas	245	Santa Ana
Hortensia	243	Santa Ana
Tierra	236	Santa Ana
Carmesí	236	Santa Ana
Rosas	176	Santa Ana
Globos	154	Santa Ana
Listón	136	Santa Ana

Producto	Cantidad
Listón	975
Rosas	945
Globos	892
Aurora	804
Orquídeas	797
Lirios	795
Girasoles	787
Tarjetas	779
Macetas	778
Hortensia	774
Tulpanes	753
Carmesí	747
Tierra	745
Claveles	733

Mejores combinaciones de productos por Departamento y País

	Departamento	Producto1	Producto2	VecesCompradasJuntas
1	San Miguel	Aurora	Rosas	90
2	San Miguel	Aurora	Orquídeas	88
3	San Miguel	Carmesí	Orquídeas	88
4	San Miguel	Aurora	Hortensia	87
5	San Miguel	Lirios	Orquídeas	86
6	San Miguel	Carmesí	Girasoles	86
7	San Miguel	Carmesí	Hortensia	85
8	San Miguel	Lirios	Rosas	85
9	San Miguel	Aurora	Carmesí	85
10	San Miguel	Rosas	Tarjetas	84
11	San Miguel	Carmesí	Rosas	83
12	San Miguel	Hortensia	Tarjetas	83
13	San Miguel	Hortensia	Rosas	83
14	San Miguel	Hortensia	Lirios	83
15	San Miguel	Hortensia	Orquídeas	83
16	San Miguel	Globos	Orquídeas	82
17	San Miguel	Carmesí	Lirios	82
18	San Miguel	Girasoles	Listón	82
19	San Miguel	Listón	Orquídeas	82
20	San Miguel	Orquídeas	Tulpanes	81
21	San Miguel	Globos	Lirios	81
22	San Miguel	Aurora	Listón	81
23	San Miguel	Carmesí	Tarjetas	80
24	San Miguel	Aurora	Tulpanes	80
25	San Miguel	Hortensia	Listón	79

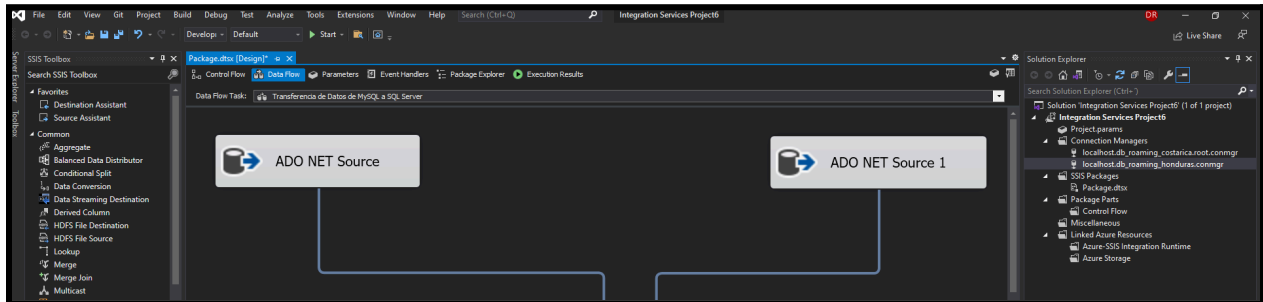
	Departamento	Producto1	Producto2	VecesCompradasJuntas
92	San Salvador	Listón	Rosas	560
93	San Salvador	Globos	Listón	540
94	San Salvador	Globos	Rosas	483
95	San Salvador	Listón	Macetas	368
96	San Salvador	Aurora	Listón	356
97	San Salvador	Listón	Orquídeas	353
98	San Salvador	Listón	Tarjetas	350
99	San Salvador	Girasoles	Listón	345
100	San Salvador	Hortensia	Listón	341
101	San Salvador	Listón	Tierra	338
102	San Salvador	Lirios	Listón	332
103	San Salvador	Macetas	Rosas	329
104	San Salvador	Claveles	Listón	324
105	San Salvador	Carmesí	Listón	320
106	San Salvador	Listón	Tulpanes	320
107	San Salvador	Rosas	Tarjetas	312
108	San Salvador	Orquídeas	Rosas	310
109	San Salvador	Globos	Macetas	310
110	San Salvador	Hortensia	Rosas	309
111	San Salvador	Aurora	Rosas	303
112	San Salvador	Girasoles	Rosas	303
113	San Salvador	Aurora	Globos	302
114	San Salvador	Globos	Hortensia	301
115	San Salvador	Globos	Orquídeas	300

	Departamento	Producto1	Producto2	VecesCompradasJuntas
183	Santa Ana	Aurora	Tarjetas	145
184	Santa Ana	Girasoles	Orquídeas	144
185	Santa Ana	Girasoles	Tarjetas	143
186	Santa Ana	Aurora	Girasoles	142
187	Santa Ana	Lirios	Orquídeas	137
188	Santa Ana	Aurora	Lirios	137
189	Santa Ana	Girasoles	Macetas	137
190	Santa Ana	Girasoles	Tulpanes	136
191	Santa Ana	Orquídeas	Tulpanes	135
192	Santa Ana	Claveles	Girasoles	134
193	Santa Ana	Girasoles	Lirios	133
194	Santa Ana	Lirios	Tarjetas	132
195	Santa Ana	Lirios	Macetas	131
196	Santa Ana	Macetas	Tarjetas	131
197	Santa Ana	Carmesí	Orquídeas	131
198	Santa Ana	Girasoles	Tierra	129
199	Santa Ana	Claveles	Hortensia	129
200	Santa Ana	Hortensia	Lirios	129
201	Santa Ana	Claveles	Lirios	128
202	Santa Ana	Girasoles	Hortensia	128
203	Santa Ana	Carmesí	Lirios	128
204	Santa Ana	Lirios	Tulpanes	127
205	Santa Ana	Aurora	Hortensia	127
206	Santa Ana	Orquídeas	Tarjetas	127
207	Santa Ana	Claveles	Macetas	126

Ejercicio #6

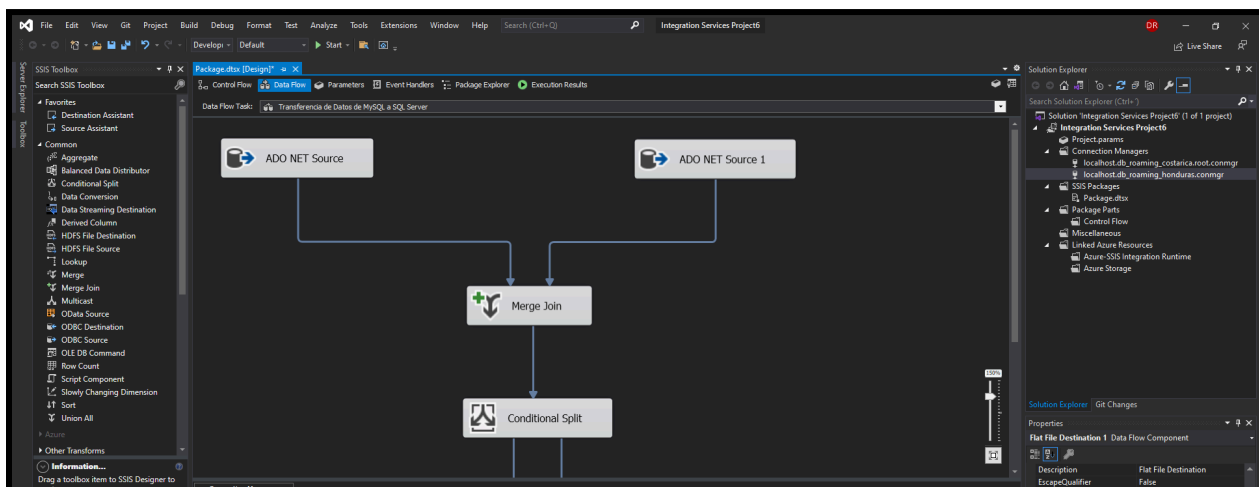
Captura de pantalla de desarrollo de ejercicio 25%

Establecimiento de conexiones a las bases de datos MySQL y SQL con la creación de los ADO NET Source en Data Flow conectados a las bases de datos realizando las extracciones, transformaciones y cargas de los datos.



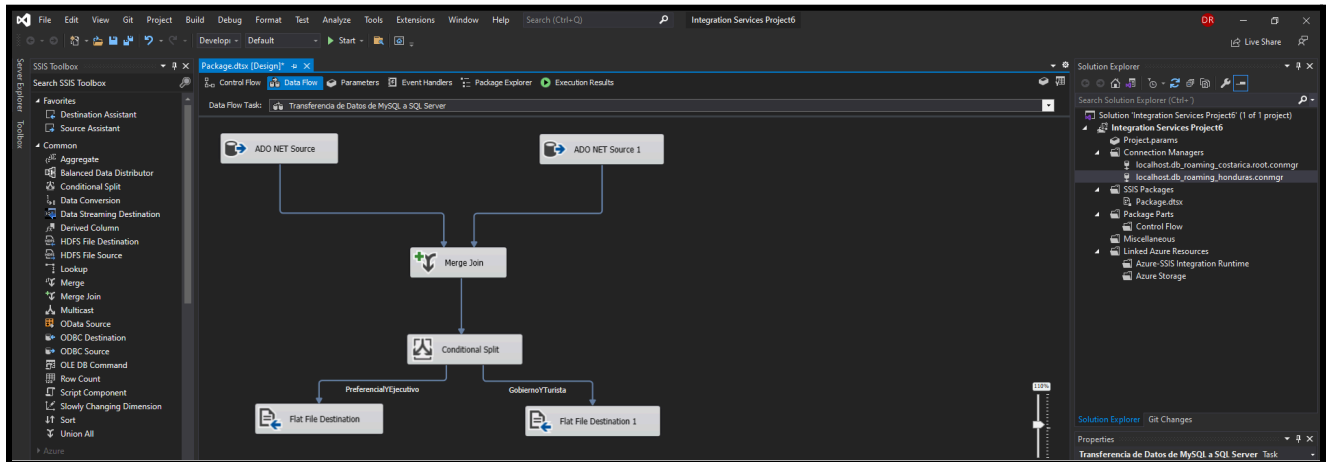
Captura de pantalla de desarrollo de ejercicio 50%

Conectar los ADO NET Source con un Merge Join y luego con el Conditional Split para separar con una condición los tipos de cliente, **ejecutivos** y **preferenciales** con los tipos de cliente **gubernamentales** y **turistas**.



Captura de pantalla de desarrollo de ejercicio 75%

Conectar el Conditional Split con un Flat File Destination que creará el archivo CSV filtrado por los tipos de cliente.



Captura de pantalla de desarrollo de ejercicio 100%

Los datos fluyen correctamente filtrados con la información que se pide de ambas carpetas de Excel.

