Experiencia de aprendizaje
Desafío 1

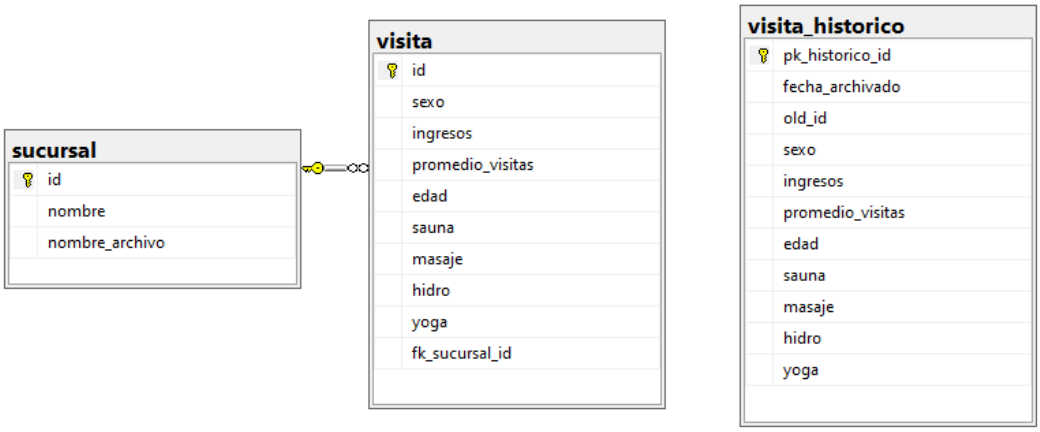
Desarrollado por **Edgar Mauricio Rivas Hernández** | RH131925

Herramientas utilizadas:

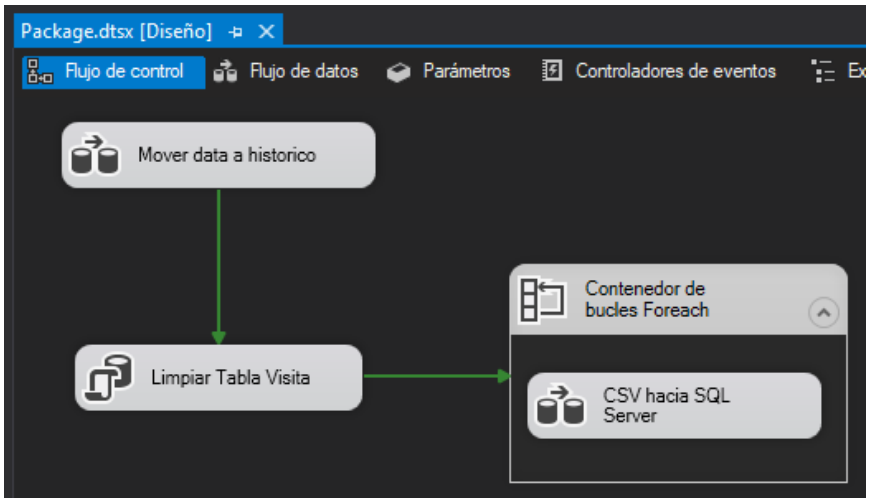
- Visual Studio 2019
- SQL Server 2019 Developer Edition
- SQL Server Management Studio 18
- SSIS v3.12

Ejercicio 1: Spa Diego

Se diseñó la siguiente base de datos para almacenar la información proporcionada por el cliente. Se separó sucursal en su propia tabla para poder utilizarla como filtro desde el ETL. Se crea una tabla visita_historico para almacenar la información actual en de la tabla visita antes de cargar nueva data.

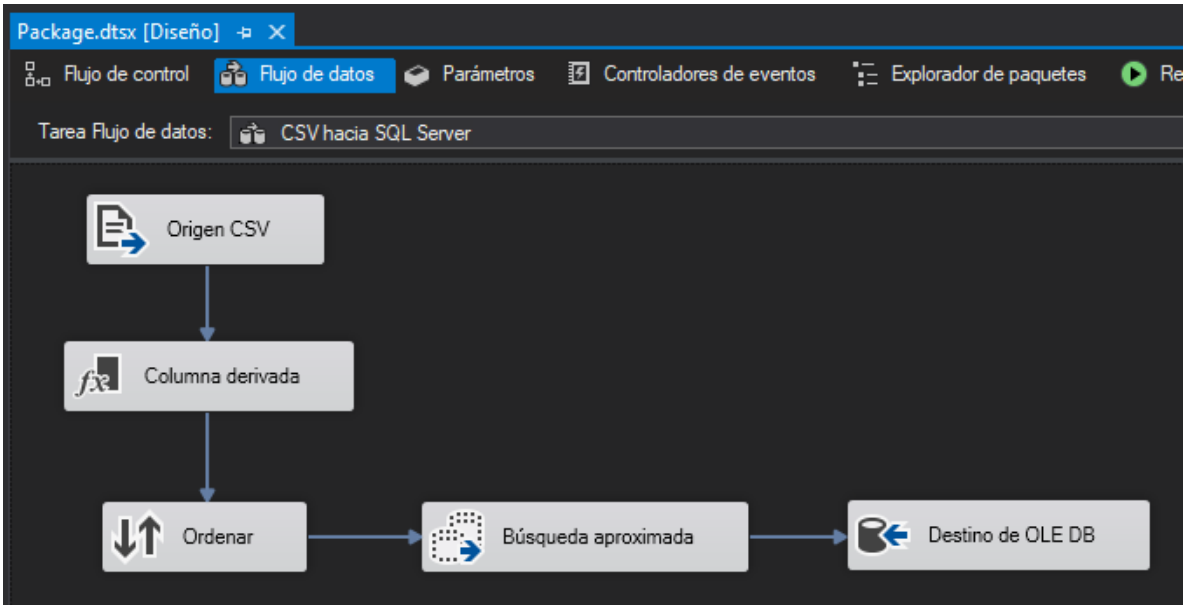


Vista general del Flujo de control: Se agrega al inicio una Tarea de flujo de datos “Mover data a histórico”, que se encarga de mover la data de visitas a visitas_historico. “Limpiar Tabla Visita” ejecuta un truncate sobre la tabla visitas. El Contenedor Foreach procesa cada archivo CSV y lo envía a la tarea de flujo principal “CSV hacia SQL Server”



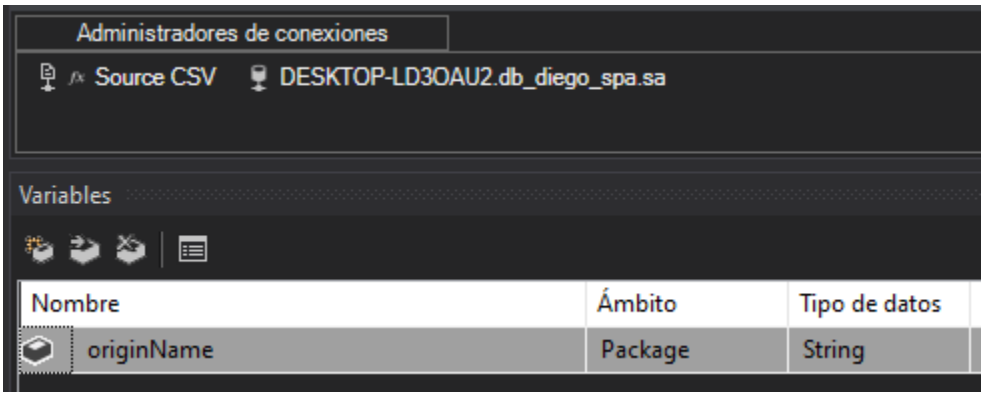
Vista del flujo “CSV hacia SQL Server”. Contiene los siguientes componentes:

1. Origen CSV, utiliza una expresión para tomar el path del archivo del ForEach mostrado anteriormente
2. Columna Derivada, agrega el nombre de archivo como columna a todos los registros
3. Ordenar, ordena los registros por sucursal
4. Búsqueda aproximada, consulta la tabla sucursales, su campo nombre_archivo para verificar a qué sucursal se debe asignar. Agrega la columna fk_sucursal_id
5. Destino de OLE DB, escribe a la base db_diego_spa

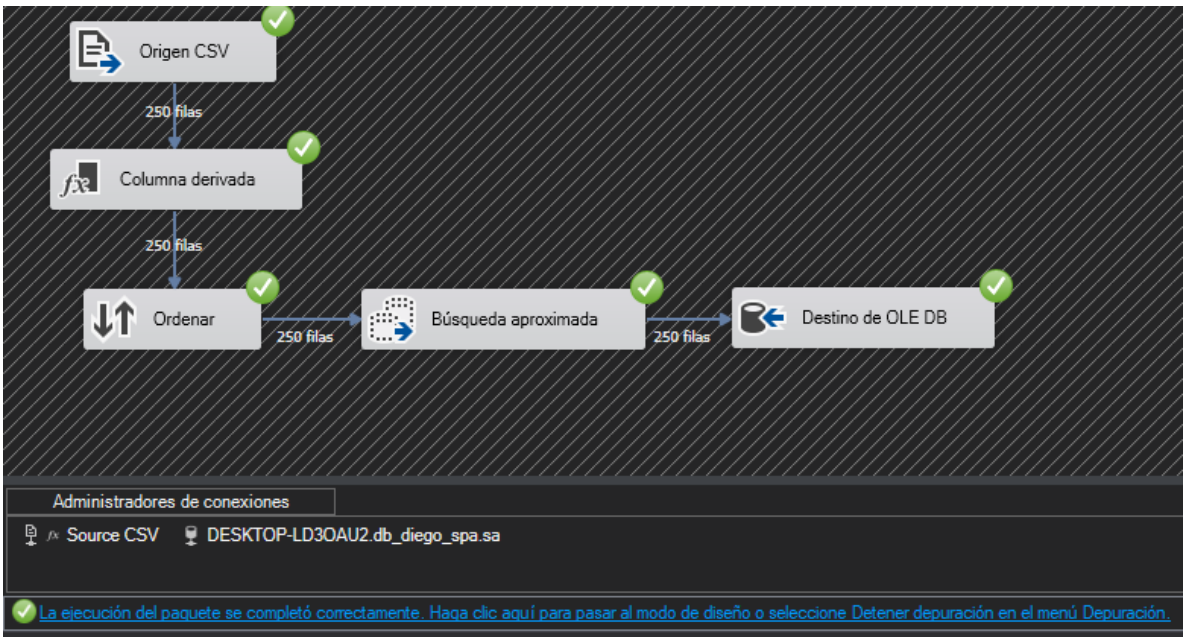


Vista de las conexiones y variables.

1. Source CSV, la fuente de archivos planos usa una expresión basada en la variable originName
2. DESKTOP-LD3OAU2.db_diego_spa.sa, la conexión al servidor SQL Server
3. originName, es la variable donde el ForEach deposita el path del archivo que se lee actualmente



Muestra del paquete en ejecución



Consultando la tabla vistas, vemos los 700 registros esperados

Results | Messages

	id	sexo	ingresos	promedio_visitas	edad	sauna	masaje	hidro	yoga	fk_sucursal_
1	Ab St Quenin	1	2678.25	4.000	31	0	0	0	1	3
2	Abbi Boyda	0	1803.34	3.000	62	1	1	0	1	3
3	Abelard Cassin	0	2077.93	6.000	38	1	1	0	1	2
4	Ad Peebles	0	1209.85	6.000	64	0	1	0	1	2
5	Addy Dillinton	0	1167.58	1.000	43	0	0	0	1	3
6	Adore Robottom	0	2723.01	4.000	61	1	1	0	0	3

Query... | DESKTOP-LD3OAU2 (15.0 RTM) | DESKTOP-LD3OAU2\PC (53) | db_diego_spa | 00:00:00 | 700 rows

Consultas para agrupar clientela.

1. Clientes / Visitas totales por sucursal
2. Clientes / Visitas totales por sexo
3. Mezcla de las anteriores

	sucursal	cantidad	visitas promedio	visitas total aprox.	Porcentaje (%)
1	Centro	50	3.4	172	7.14
2	Santa Tecla	250	3.3	835	35.71
3	Escalón	400	3.5	1398	57.14

	sexo	cantidad	visitas promedio	visitas total aprox.	Porcentaje (%)
1	Mujeres	304	3.4	1041	43.43
2	Hombres	396	3.4	1362	56.57

	sucursal	sexo	cantidad	visitas promedio	visitas total aprox.	Porcentaje (%)
1	Centro	Hombres	14	3.6	50	2.00
2	Centro	Mujeres	36	3.4	120	5.14
3	Santa Tecla	Mujeres	55	3.4	184	7.86
4	Escalón	Hombres	187	3.5	661	26.71
5	Santa Tecla	Hombres	195	3.3	649	27.86
6	Escalón	Mujeres	213	3.5	735	30.43

4. Por rango de ingresos, mostrando ingresos promedio.

Results Messages

	rango_ingresos	sexo	cantidad	visitas promedio	visitas total aprox.	media_ingresos
1	0 - 500	Mujer	21	3.7	77	355.4309
2	0 - 500	Hombre	32	3.3	106	356.9021
3	500 - 1000	Mujer	50	3.3	163	774.6524
4	500 - 1000	Hombre	69	3.0	205	763.6662
5	1000 - 2000	Mujer	114	3.4	392	1560.5155
6	2000 - 3000	Mujer	119	3.4	407	2491.4776
7	2000 - 3000	Hombre	136	3.7	499	2558.8113
8	1000 - 2000	Hombre	159	3.5	550	1529.2261

Se puede detectar con esto los siguientes puntos a considerar:

1. Hay más clientela masculina
2. Los clientes que ganan más de $1000 conforman casi dos terceras partes de la clientela
3. La sucursal más visitada es Escalón, en contraste la sucursal Centro casi no tiene visitas
4. Las mujeres que ganan menos de $500 conforman el grupo más pequeño de la clientela (21), y los hombres que ganan entre $1000 y $2000 son el grupo más grande (159)

Otra consulta de interés es el uso de servicios, podemos ver que el servicio más usado por si solo es Yoga. Y el que menos clientes usan es el Sauna.

Results | Messages

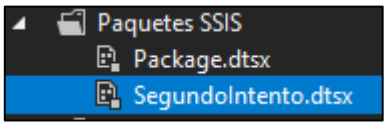
	Paquete	visitas total aprox.
1	Sauna e Hidro	60
2	Completo (4 Servicios)	93
3	Sauna	102
4	Hidro	112
5	Hidro y Yoga	113
6	Masaje e Hidro	124
7	Sauna y Yoga	126
8	Masaje	189
9	Sauna y Masaje	195
10	Yoga	235
11	Masaje y Yoga	259

Ejercicio 2

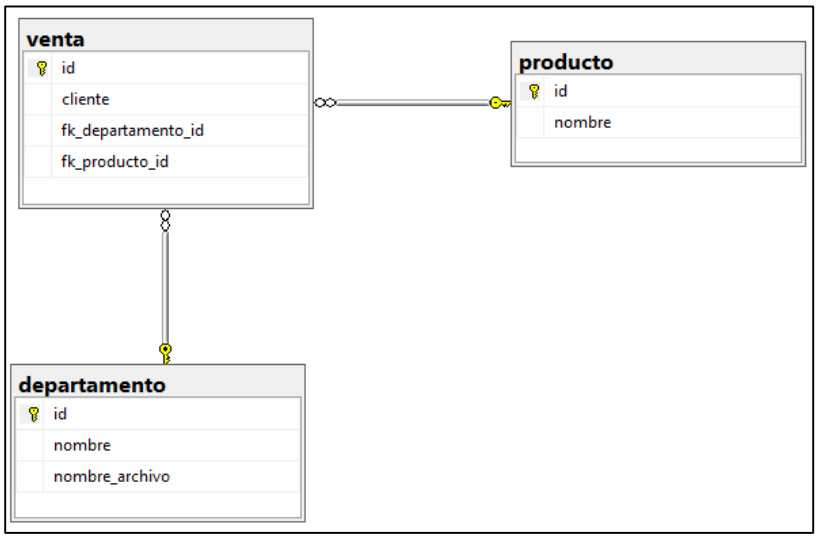
Nota: este ejercicio contiene dos soluciones, inicialmente se planteó crear una tabla con todas sus dimensiones incluidas, excepto por el departamento, pero esta versión resultó inviable para realizar consultas. Opté por realizar un segundo intento agregando la dimensión “producto” para facilitar el análisis de los resultados.

Todos los archivos del primer intento se encuentran dentro de la carpeta Ejercicio2

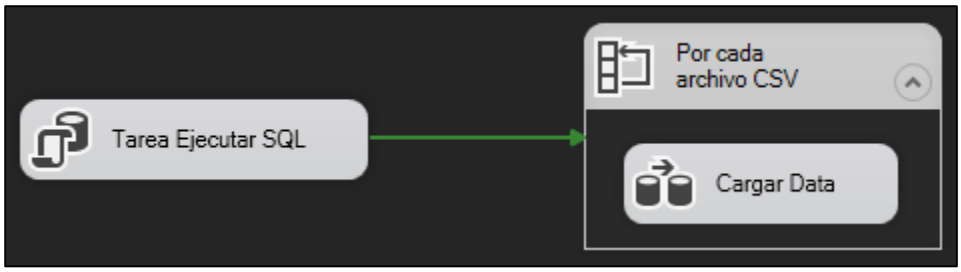
Desarrollo de segundo intento



Para este ejercicio desarrollé la siguiente base de datos, donde la tabla de hecho es “venta”. Las tablas “producto” y “departamento” corresponden a dimensiones de la venta.

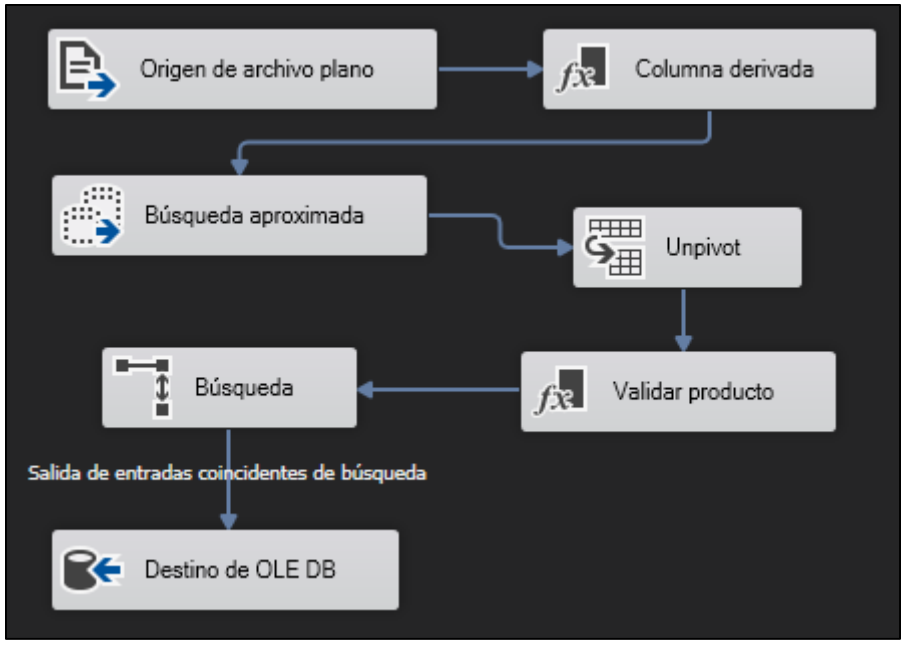


El flujo de control es el siguiente, iniciando por un truncado de la tabla venta, para ingresar la nueva data. Un contenedor ForEach para cada archivo CSV y el Flujo de datos “Cargar data”



El flujo de datos contiene estos componentes:

1. Origen de archivo plano para los archivos CSV
2. Columna derivada para agregar el path de archivo como columna
3. Búsqueda aproximada para hacer match entre el nombre de archivo en base de datos y el path almacenado en la columna en el paso anterior
4. Anulación de dinamización “Unpivot”, sirve para convertir todas las columnas de producto en registros de la tabla venta, a partir de aquí se hace un loop a través de cada columna afectada
5. Columna derivada para agregar columna “producto” en la cual se guarda el nombre de la columna a procesar en ese momento, ejemplo producto=Rosas
6. Búsqueda, para hacer match exacto entre el nombre de la columna a procesar y el producto guardado en la tabla “producto”
7. Guardar un registro en ventas por cada columna de producto a procesar.



El resultado en la base de datos es el siguiente (consulta: select * from [dbo].[venta] v inner join [dbo].[producto] p on p.id = v.fk_producto_id)

Results | Messages

	id	cliente	fk_departamento_id	fk_producto_id	id	nombre
1	4	Egon Greenhead	1	5	5	Girasoles
2	5	Egon Greenhead	1	7	7	Globos
3	6	Egon Greenhead	1	6	6	Hortensia
4	7	Egon Greenhead	1	11	11	Lirios
5	8	Egon Greenhead	1	14	14	Liston
6	11	Egon Greenhead	1	1	1	Rosas
7	16	Elita Borles	1	10	10	Carnesi

Query... | DESKTOP-LD3OAU2 (15.0 RTM) | DESKTOP-LD3OAU2\PC (53) | second_try | 00:00:00 | 11,304 rows

Utilizando una consulta recursiva se obtiene todas las combinaciones de productos y la cantidad de veces que se repite dicha combinación, como se puede observar, la cantidad de combinaciones hace inservible el resultado. Hablamos de poco menos de 42 mil combinaciones.

	departamento	productos_diferentes	combinacion	cantidad
166	Santa Ana	3	Claveles - Macetas - Lirios	63
167	Santa Ana	3	Claveles - Orquidias - Tulipanes	63
168	Santa Ana	3	Claveles - Tarjetas - Aurora	63
169	Santa Ana	3	Claveles - Macetas - Hortensia	62
170	Santa Ana	3	Carmesi - Lirios - Tulipanes	62
171	Santa Ana	3	Tierra - Girasoles - Carmesi	62
172	Santa Ana	3	Tierra - Hortensia - Aurora	62

Query... | DESKTOP-LD3OAU2 (15.0 RTM) | DESKTOP-LD3OAU2\PC (53) | second_try | 00:00:29 | 41,825 rows

Para limitar este resultado y volverlo útil se utiliza algunas consultas para determinar algunos datos de interés

1. Top 3 productos en por departamento

Results | Messages

	departamento	veces comprado	producto
1	San Miguel	160	Aurora
2	San Miguel	160	Lirios
3	San Miguel	158	Orquidias

	departamento	veces comprado	producto
1	San Salvador	690	Liston
2	San Salvador	612	Rosas
3	San Salvador	587	Globos

	departamento	veces comprado	producto
1	Santa Ana	270	Lirios
2	Santa Ana	266	Girasoles
3	Santa Ana	260	Aurora

2. Top 5 combinaciones por departamento, se observa una tendencia a comprar por separado

Results | Messages

	departamento	productos_diferentes	combinacion	cantidad
1	San Salvador	1	Liston	690
2	San Salvador	1	Rosas	612
3	San Salvador	1	Globos	587
4	San Salvador	2	Rosas - Liston	560
5	San Salvador	2	Globos - Liston	540

Results | Messages

	departamento	productos_diferentes	combinacion	cantidad
1	San Miguel	1	Aurora	160
2	San Miguel	1	Lirios	160
3	San Miguel	1	Carmesi	158
4	San Miguel	1	Orquidias	158
5	San Miguel	1	Hortensia	157

Results | Messages

	departamento	productos_diferentes	combinacion	cantidad
1	Santa Ana	1	Lirios	270
2	Santa Ana	1	Girasoles	266
3	Santa Ana	1	Aurora	260
4	Santa Ana	1	Orquidias	259
5	Santa Ana	1	Tarjetas	252

3. Consultar cantidades de repeticiones por número de productos diferentes para saber cuántos productos compra un cliente en promedio, muestra de San Salvador como referencia. Encontramos un dato de mucho interés, en los 3 departamentos, aunque no en el mismo orden, las ventas incluyen más veces 3, 4, 5 y 6 productos. Nos podemos enfocar en ese detalle para el último análisis

Santa Ana	4	21452
Santa Ana	5	19265
Santa Ana	3	17322
Santa Ana	6	12951
Santa Ana	2	9567
Santa Ana	7	6629
Santa Ana	1	3226

San Salvador	5	86594
San Salvador	4	76968
San Salvador	6	72848
San Salvador	3	49643
San Salvador	7	46473
San Salvador	8	22522
San Salvador	2	21967

San Miguel	5	19129
San Miguel	4	19058
San Miguel	6	14483
San Miguel	3	13825
San Miguel	7	8468
San Miguel	2	6890
San Miguel	8	3883

Análisis final Ejercicio 2

Para entregar un informe a la floristería Fiorella se puede consultar por departamento, las combinaciones más usadas para 3, 4, 5 y 6 productos. Ejemplos

1. Top 3 combinaciones de 3 productos para Santa Ana

Results | Messages

	departamento	productos_diferentes	combinacion	cantidad
1	Santa Ana	3	Girasoles - Tarjetas - Aurora	83
2	Santa Ana	3	Macetas - Girasoles - Tarjetas	78
3	Santa Ana	3	Claveles - Girasoles - Tarjetas	74

2. Top 10 combinaciones de 6 productos para San Miguel

Results | Messages

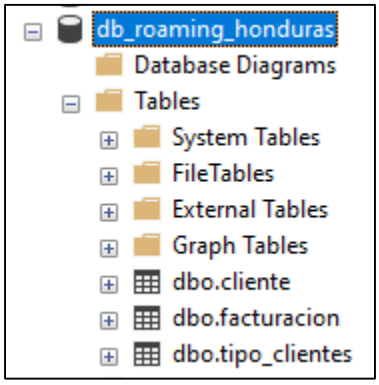
	departamento	productos_diferentes	combinacion	cantidad
1	San Miguel	6	Claveles - Globos - Orquidias - Lirios - Tulipanes - Liston	12
2	San Miguel	6	Rosas - Macetas - Hortensia - Tarjetas - Carmesi - Aurora	12
3	San Miguel	6	Claveles - Globos - Orquidias - Carmesi - Lirios - Liston	11
4	San Miguel	6	Claveles - Globos - Orquidias - Carmesi - Lirios - Tulipanes	11
5	San Miguel	6	Claveles - Hortensia - Globos - Orquidias - Aurora - Liston	11
6	San Miguel	6	Macetas - Hortensia - Tarjetas - Orquidias - Lirios - Aurora	11
7	San Miguel	6	Rosas - Claveles - Tarjetas - Orquidias - Aurora - Tulipanes	11
8	San Miguel	6	Tierra - Girasoles - Globos - Orquidias - Tulipanes - Liston	11
9	San Miguel	6	Claveles - Globos - Carmesi - Lirios - Tulipanes - Liston	10
10	San Miguel	6	Claveles - Globos - Orquidias - Carmesi - Tulipanes - Liston	10

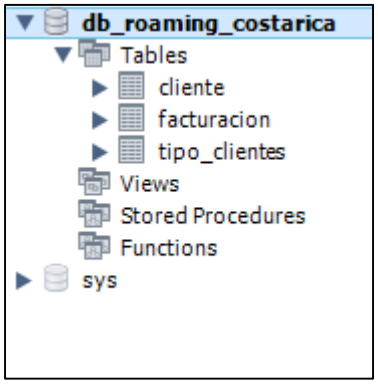
Conclusión

Con la base de datos y el proceso de integración de datos que se ha creado se puede realizar consultas muy sencillas con las cuales se determine la cantidad de productos que se consume por departamento y también por sus posibles combinaciones.

Ejercicio 3

Para el ejercicio 3 se creó ambas bases de datos en sus respectivos motores de base de datos SQL Server y MySQL

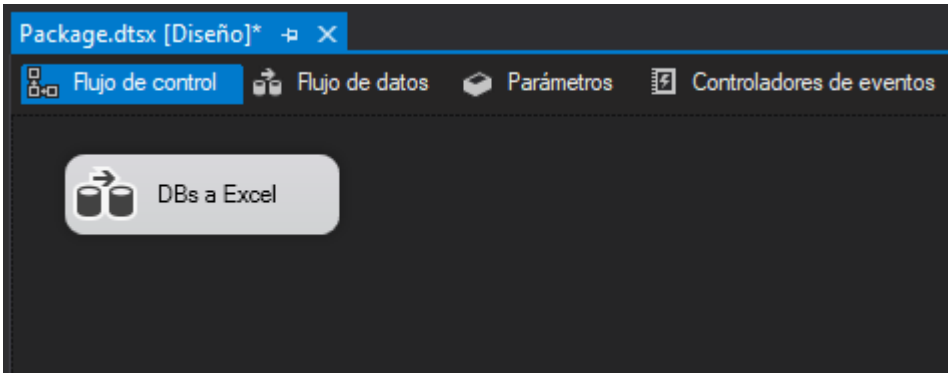




Se creó 2 archivos Excel llamados “Preferencial y Ejecutivo.xlsx” y “Gobierno y Turista.xlsx” con el siguiente formato como destino de la data en las bases de datos

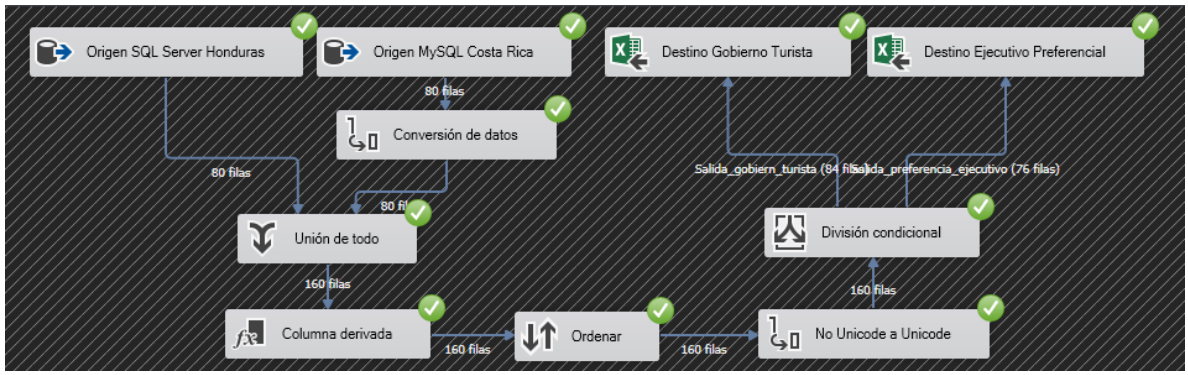
	A	B	C	D	E	F	G	H	I	J
1	COD PAÍS	COD CLIENTE	NOMBRES	APELLIDOS	DUI	NIT	SEXO	TELÉFONO	FACTURACIÓN ($)	ESTADO
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

El Flujo de control es muy básico, solamente incluye la siguiente Tarea de flujo de datos



El Flujo de datos incluye los siguientes componentes

1. Origen de base SQL Server para Honduras
2. Origen de base MySQL para Costa Rica
3. Conversión de datos, cualquiera de los dos orígenes puede ser seleccionado para este paso, pero se debe garantizar que los tipos de datos son iguales en ambas antes de poder unirlas.
4. Unión de todo
5. Columna Derivada, basado en los códigos de empleado “sv-xx00000” se intenta extraer los valores “xx” con la expresión rigth(left(codigo_empleado, 5), 2)
6. Ordenar por tipo de cliente
7. No Unicode a Unicode, ya que no todos los data types se corresponden a Unicode que espera el destino de Excel
8. División condicional, para separar los Gubernamentales y Turistas de los Ejecutivos y Preferenciales
9. Destino Gobierno Turista
10. Destino Ejecutivo Preferencial



Resultado en los archivos

	A	B	C	D	E	F	G	H	I	J
1	COD PAÍS	COD CLIENTE	NOMBRES	APELLIDOS	DUI	NIT	SEXO	TELÉFONO	FACTURACIÓN ($)	ESTADO
2	lj	sv-lj47675	JORGE ALBERTO	LOPEZ ROMERO	034541664	08212801801052	m	71146362	59.20	t
3	cm	sv-cm1664	MARTA ALICIA	CABRERA MARTINEZ	015965532	02100911630058	f	62805523	101.40	t
4	dm	sv-dm47697	MANUEL HERIBERTO	DURAN RODRIGUEZ	007291791	06091504570018	m	78804336	176.20	f
5	gm	sv-gm47660	MANUEL ENRIQUE	GRANDE CASTELLANOS	011299272	08051205721013	m	62845726	93.10	f
6	jj	sv-jj47659	JOSE RODOLFO	JIMENEZ PINTIN	005574868	02032303831017	m	62634944	26.00	t
7	nj	sv-nj26082	JOSE ANTONIO	NOLASCO LOPEZ	004813112	07032503651011	m	77109443	109.90	t
8	hm	sv-hm47684	MILTON SAMUEL	HERNANDEZ CANTARELY	044322193	06141203911305	m	75812813	96.20	f
9	cc	sv-cc47631	CARLOS ADALBERTO	CASTRO AGUILAR	022169985	01122605741012	m	65414440	42.80	t
10	ae	sv-ae47669	ESTANISLAO	ALVARADO ALVAREZ	004536087	08113011611018	m	79939305	44.10	t
11	le	sv-le47652	ERIC LOMBARDO	LEMUS ESCALANTE	017998303	06142103701059	m	62849154	16.40	t
12	me	sv-me47714	ERIC ALEXANDER	MEJIA MOLINA	004495689	10091002791017	m	70831594	30.70	t
13	ae	sv-ae27200	ELMER ENRIQUE	AREVALO	022220448	07110810661017	m	78533506	180.20	t
14	cc	sv-cc47631	CARLOS ADALBERTO	CASTRO AGUILAR	022169985	01122605741012	m	65414440	42.80	t
15	ra	sv-ra2158	ANA LIZ	RODRIGUEZ DE TOVAR	002873946	02102501671058	f	64405701	185.10	f
16	on	sv-on16424	NATIVIDAD CRISTINO	ORELLANA CHICA	029324227	13072512641016	f	62061046	60.90	t
17	sh	sv-sh23204	HECTOR ALEJANDRO	SORIANO BONILLA	037696573	06142707871033	m	77953576	102.60	t
18	ap	sv-ap47661	PEDRO ALFONSO	AMAYA ORTEGA	034079691	06090612851027	m	73910697	154.00	t

Como se puede ver se encuentran los datos ordenados y la columna COD PAÍS es llenada por el código diferente dentro de COD CLIENTE, entre sv- y el número de cliente