Primer Desafío Practico

Datawarehouse y minería de datos

Universidad Don Bosco

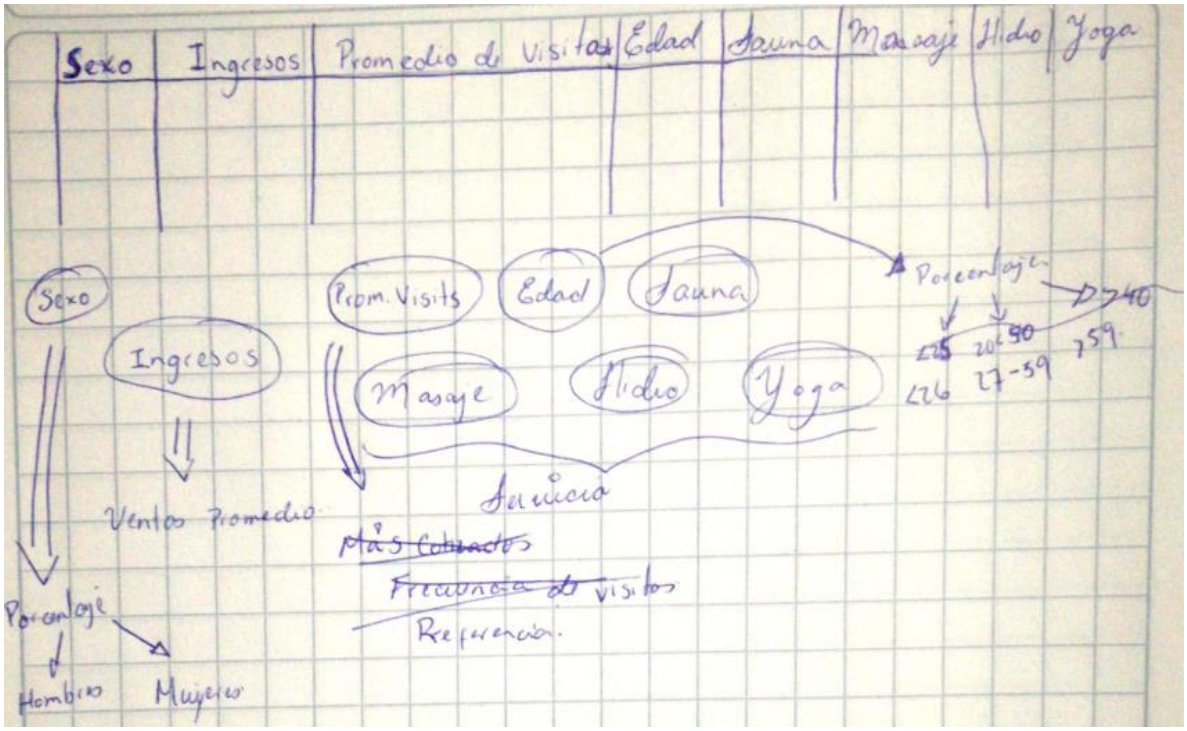
Cruz González, José Roberto	CG181933
Garay Alvarado, Bryan Walberto	GA181935

DESARROLLO

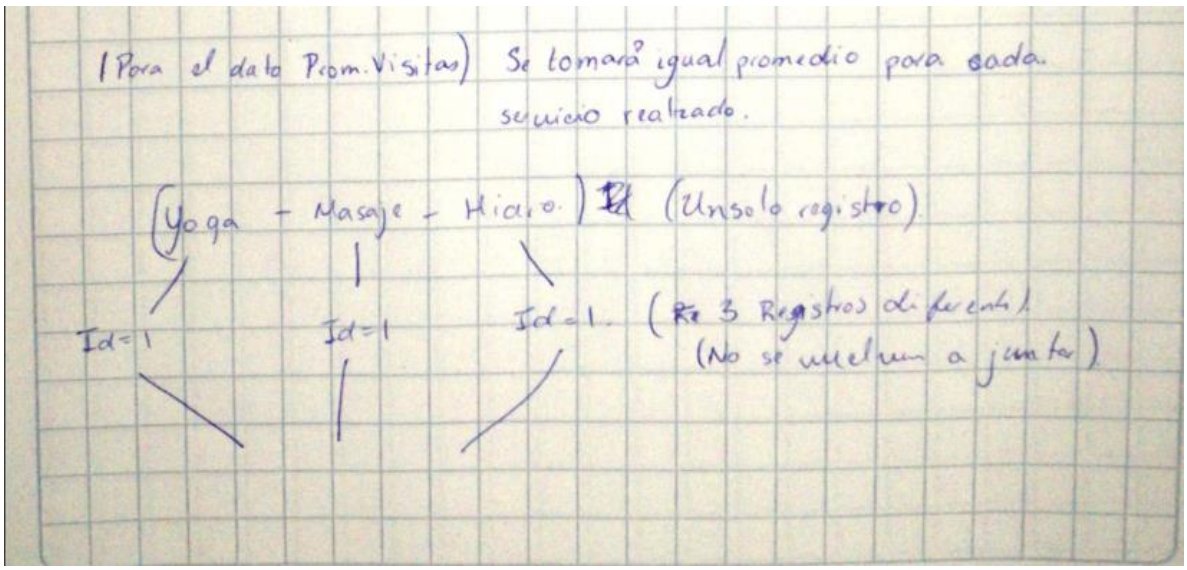
Ejercicio 1

Análisis de datos y flujo de datos

Planteamiento del ordenamiento de datos para su correspondiente análisis (en la solución se explica el por qué y las conclusiones de este planteamiento)



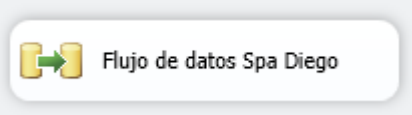
Luego se declara el sentido de los datos que son un poco dudosos acerca del sentido que cobran en la tabla:



Además, se muestra un poco de las inconsistencias que pueden surgir del planteamiento de flujo de datos.

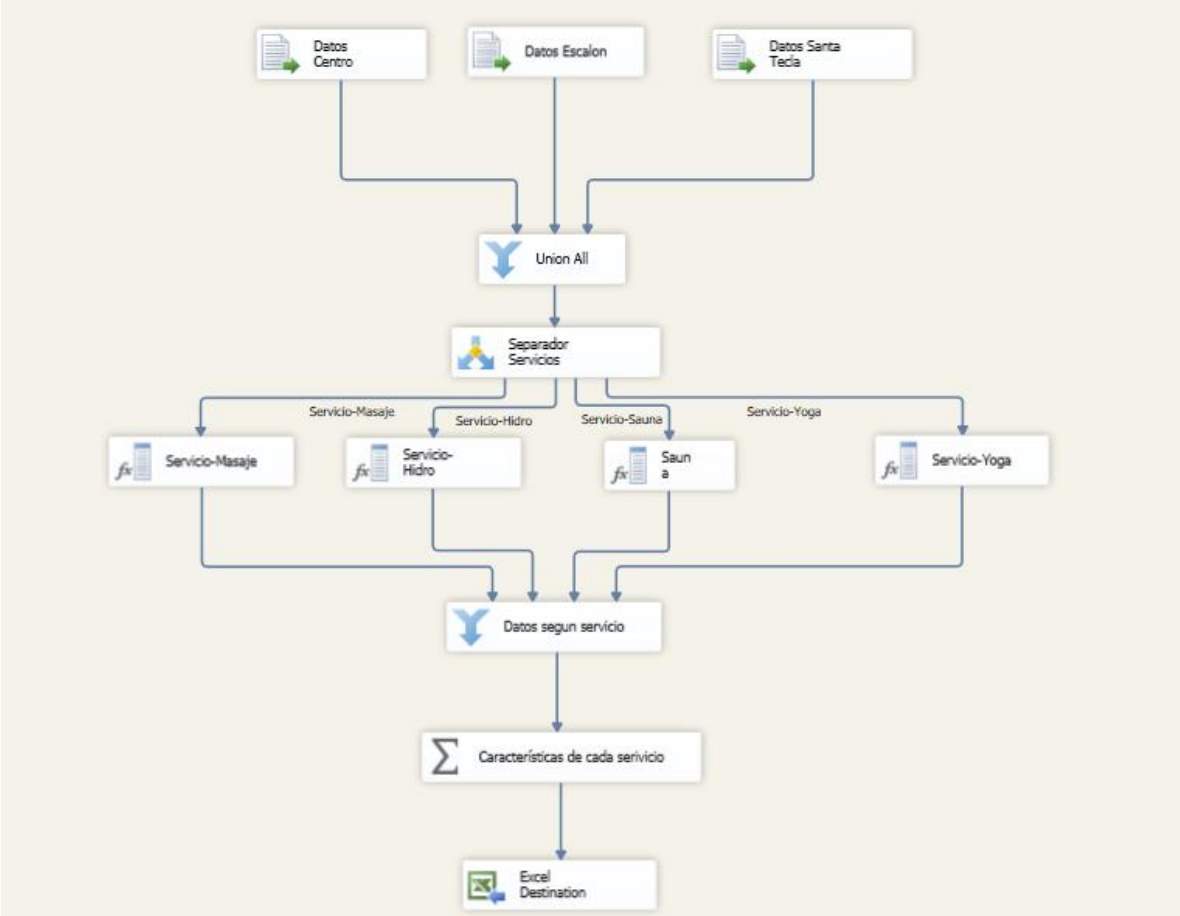
Funcionamiento y resultado

Porcentaje de realización: 100%



En el panel de control de flujo se añadió una tarea, que es donde está diseñado el ETL que obtiene los datos y realiza los procesos para obtener la información requerida.

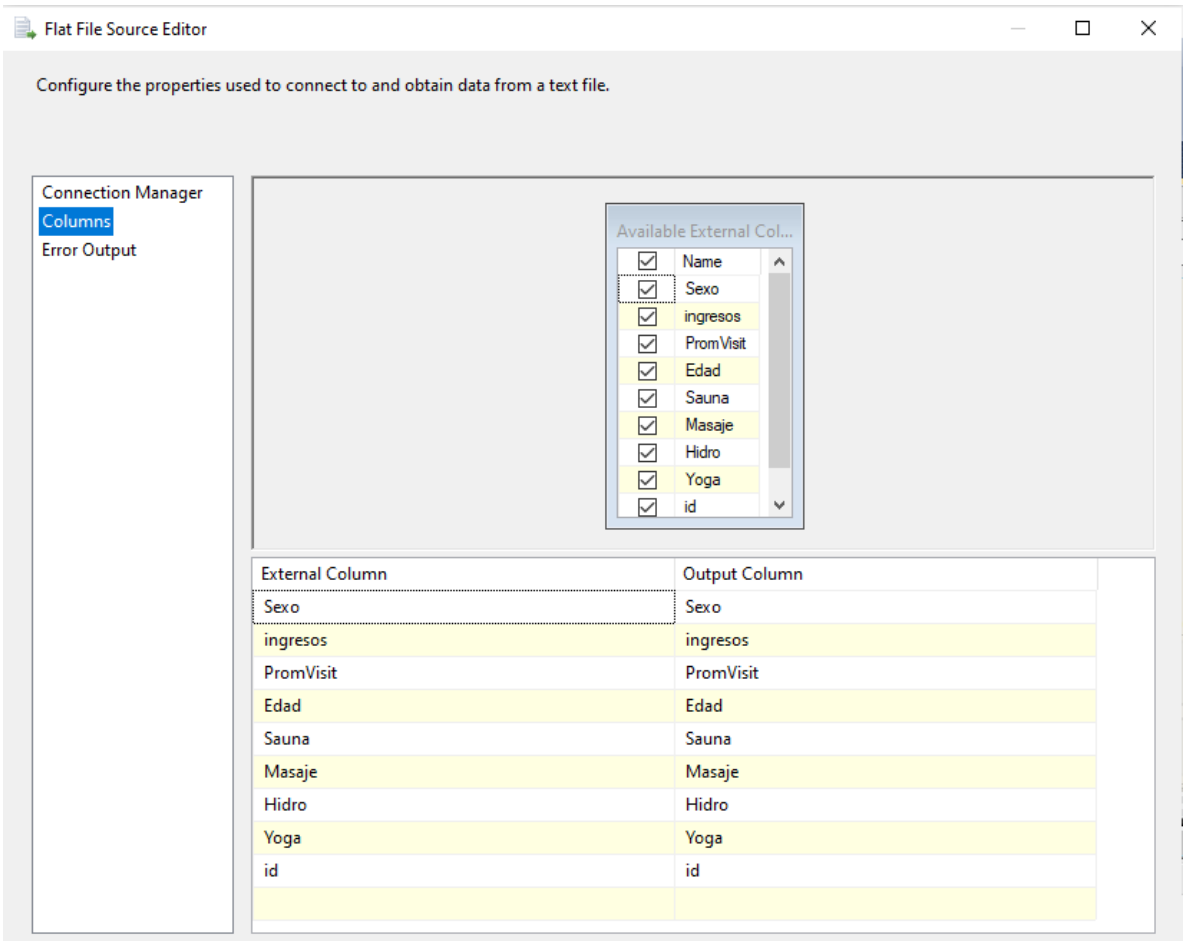
El flujo completo de datos está comprendido por este diseño:



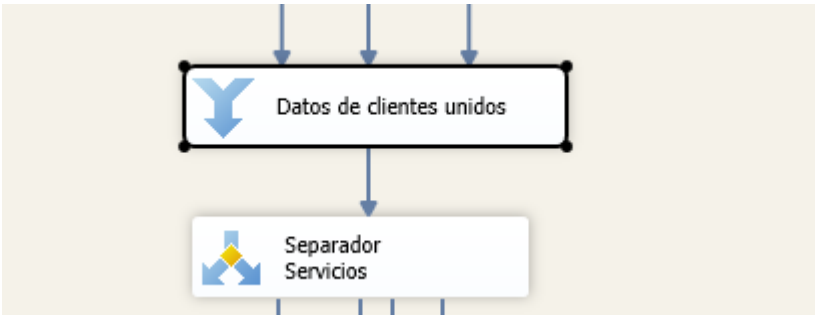
Obtención de datos:

Para la obtención de datos se realiza por medio de los controles: ***"Flat File Source"***, porque a pesar de que es un archivo de Excel, todos los IDE para este manejo de datos los toman como *"archivos delimitados"*, y en específico estos archivos se encuentran delimitados por *"comas"*, y por *"saltos de línea"*.

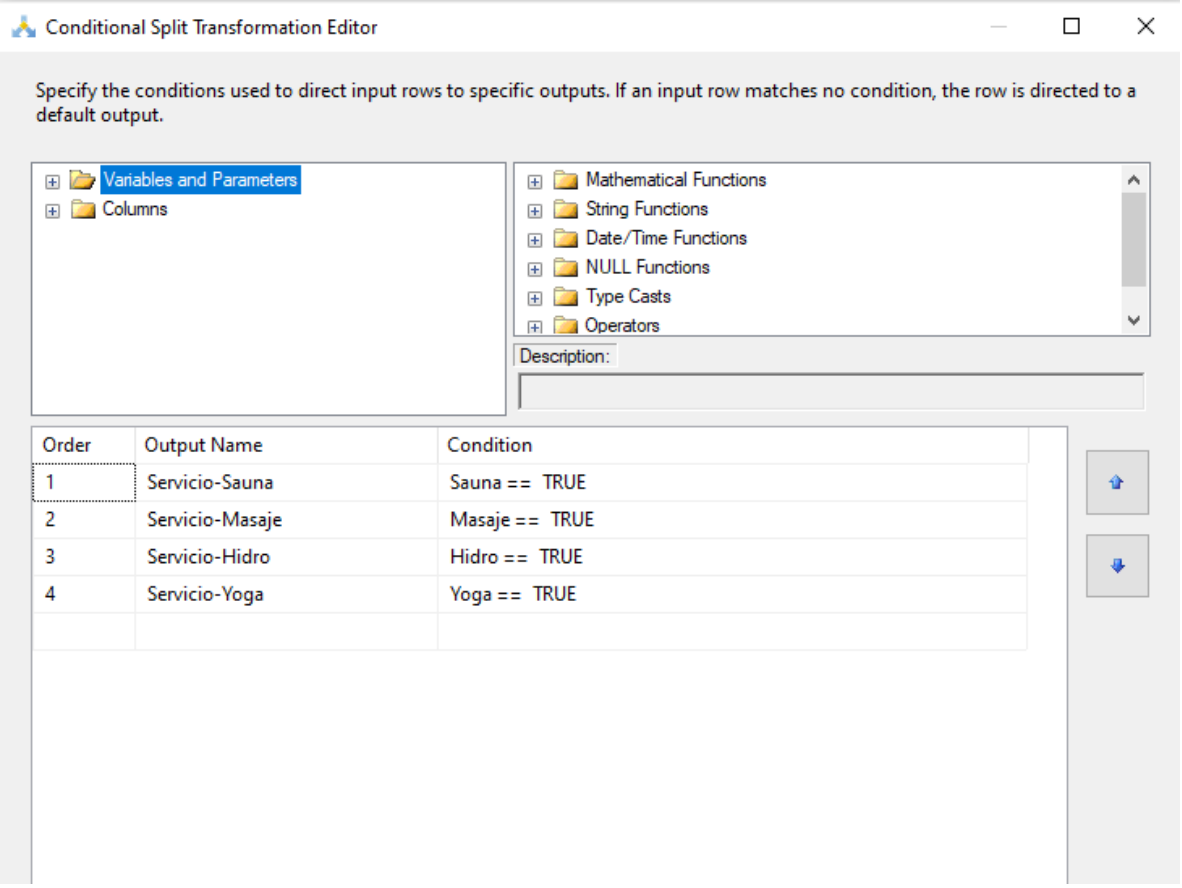
La configuración de cada conexión obtiene y establece las columnas de la siguiente manera:



Luego de esto, los datos se unen para juntar datos de id's de personas repetidas (por si ha visitado o tiene registro en diferentes sucursales) posteriormente estos se dividen nuevamente, pero ahora por tipo de servicio, es decir ahora sin separación de sucursal diferente, sino por servicio brindado en todas las sucursales:

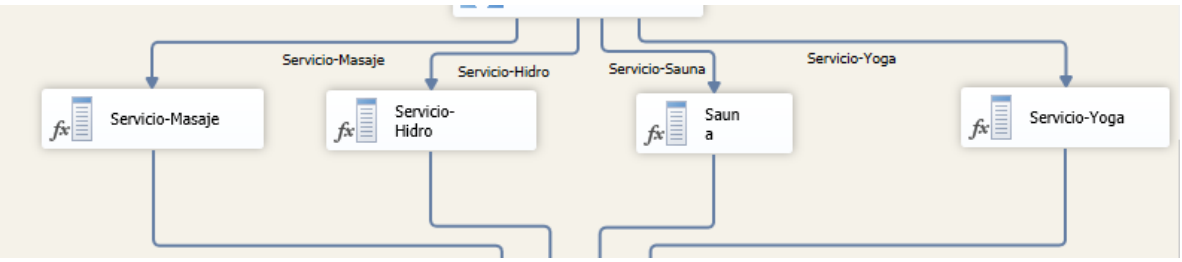


El Split condicional se realiza de la siguiente manera, viendo cuando un registro tenga en valor **TRUE**, cualquiera de los diferentes servicios, y si se visitan por ejemplo 3 al mismo tiempo este registro se divide ahora 3 diferentes, categorizándolos por el servicio:

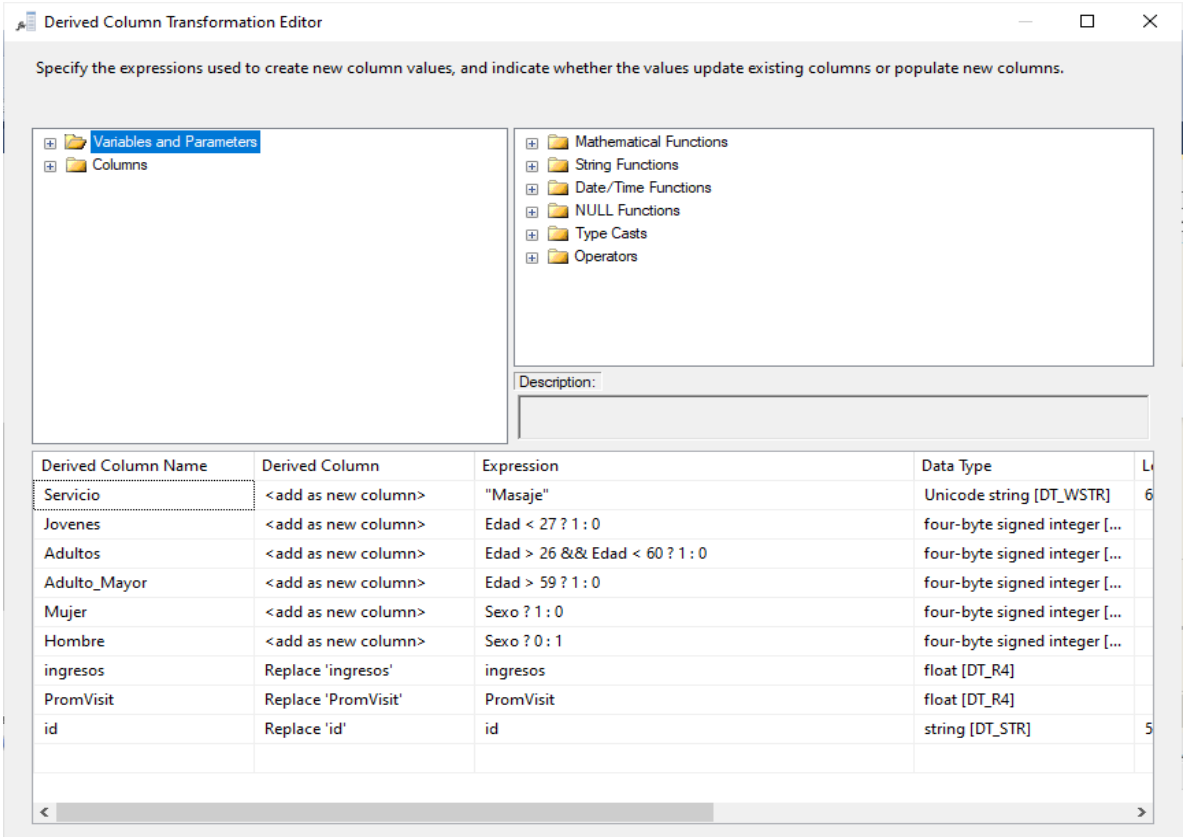


Luego de la obtención de los registros categorizados por el servicio que se da a cada cliente, pasan a ser transformados en una tabla derivada, para transformar los datos para luego recopilarlos en el final, se puede observar como hasta ahora la estructura ha sido la siguiente:

Carga de datos -> Unión de datos -> Categorización de datos unidos -> Transformación de datos



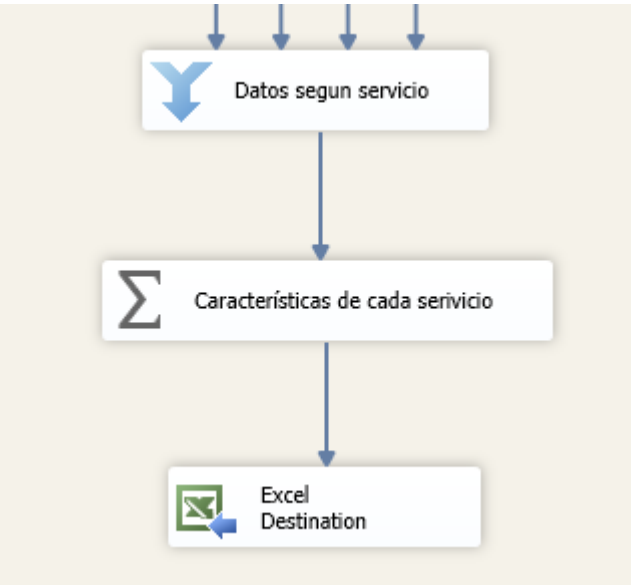
Los datos transformados por cada servicio son los siguientes:



Las columnas pasan a ser las siguientes:

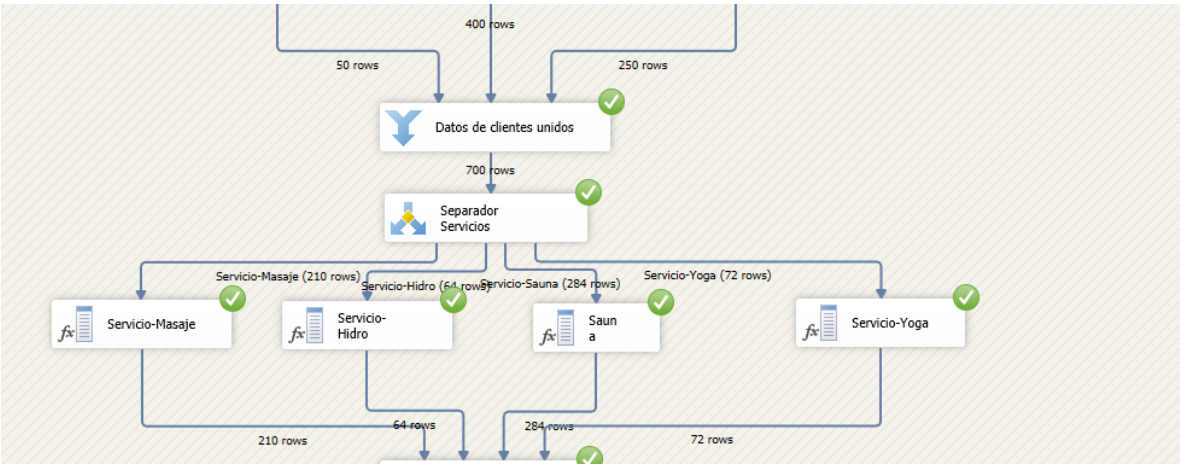
Columnas	
Servicio	Indica el tipo de servicio (como está separado es constante para cada tabla derivada)
Jóvenes	Indica si el cliente es menor de 27 años (Si se cumple el valor tomado será de 1)
Adultos	Indica si el ciente tiene una edad entre 26 y 60 años (Si se cumple el valor tomado será de 1)
Adulto_Mayor	Indica si el cliente tiene una edad mayor a 59 años (Si se cumple el valor tomado será de 1)
Mujer	Indica si el sexo es TRUE (Hemos designado **TRUE** como mujer por los nombres que hemos visto en los datos) si el valor es true se le indica el valor de 1, véase que el tipo de dato no es para indicar verdadero sino como contador entero.
Hombre	Indica si el sexo es FALSE (Hombre) si el valor es false se le indica el valor de 1, véase que el tipo de dato no es para indicar verdadero sino como contador entero.
Ingresos	Valor replicado de Ingresos
PromVisit	Valor replicado de PromVisit
Id	Nombre de cliente

Y por último se unen los datos, para luego sumarlos y colocarlos en el archivo destino:

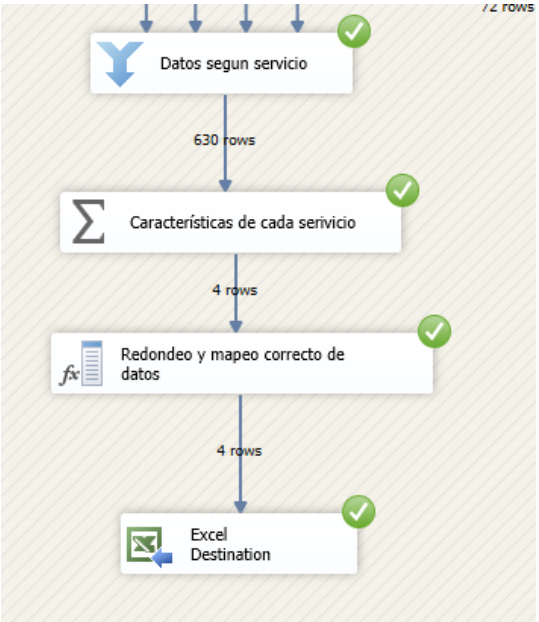


La sumatoria es para agrupar por cada servicio diferente, y obtener de cada uno datos como qué grupo de personas (Joven, Adulto o adulto mayor) consume más qué tipo de servicio, además también qué sexo prefiere más qué servicio.
Por eso se planteó de esa manera.

Puede verse el resultado de la ejecución del diseño que es satisfactorio:



Y se finaliza con 4 filas, que son los 4 servicios diferentes del spa:



Se puede observar que se ha añadido un control de columna derivada, esto no modifica la estructura de los datos, sino que solamente ajusta los valores para realizar un buen mapeo para que Excel lo reconozca de mejor manera, y también

Derived Column Name	Derived Column	Expression
Servicio	Replace 'Servicio'	Servicio
ingresos	Replace 'ingresos'	ROUND(ingresos,2)
PromVisit	Replace 'PromVisit'	ROUND(PromVisit,2)
Jovenes	Replace 'Jovenes'	ROUND(Jovenes,4)
Adultos	Replace 'Adultos'	ROUND(Adultos,4)
Adulto_Mayor	Replace 'Adulto_Mayor'	ROUND(Adulto_Mayor,4)
Hombre	Replace 'Hombre'	ROUND(Hombre,4)
Mujer	Replace 'Mujer'	ROUND(Mujer,4)

realiza aproximaciones para mejor lectura de los datos de la siguiente manera:

El resultado de los datos es el siguiente:

A	B	C	D	E	F	G	H
Servicio	ingresos	PromVisit	Jovenes	Adultos	Adulto_Mayor	Hombre	Mujer
Masaje	1710.86	3.92	0.1287	0.7571	0.1143	0.6143	0.3857
Hidro	1738.35	4	0.1562	0.7656	7.81E-02	0.5156	0.4844
Yoga	1692.75	3.78	0.1944	0.6389	0.1667	0.6806	0.3194
Sauna	1689.8	3.99	0.1232	0.757	0.1197	0.507	0.493

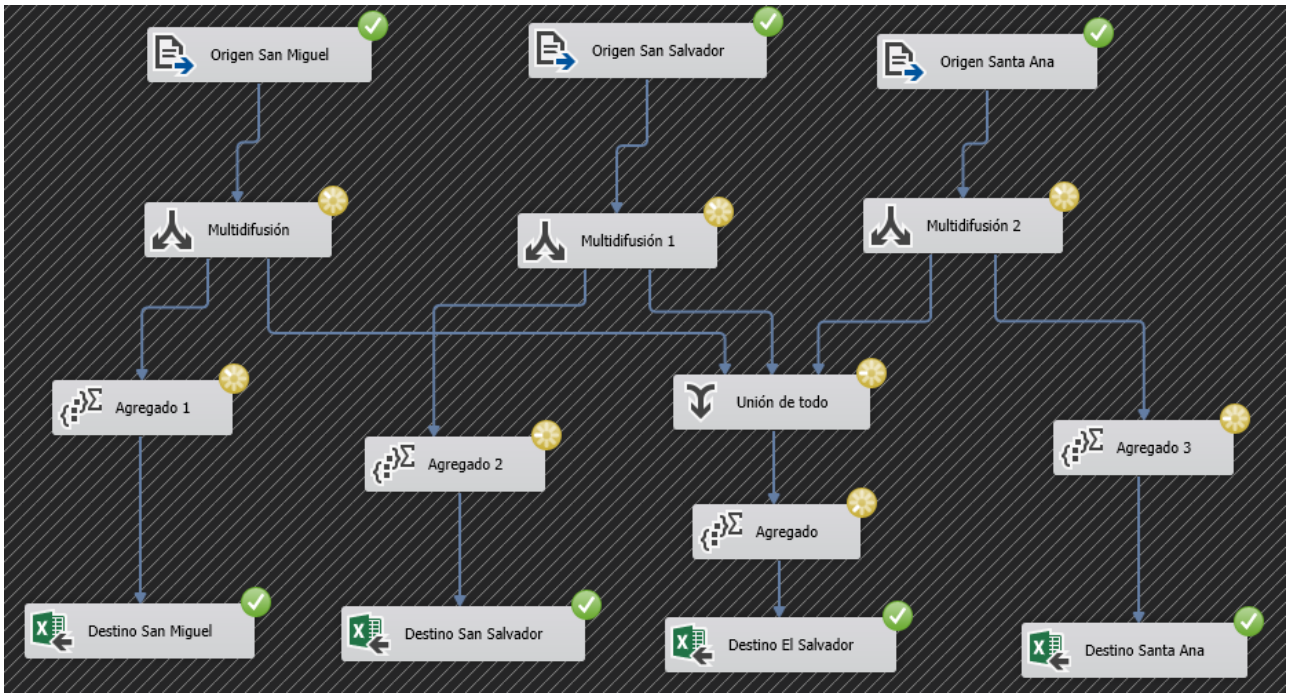
Podemos concluir de esta estructura, por ejemplo, que el servicio más cotizado es el Hidro, donde la mayor cantidad de personas son adultos.

Otro ejemplo, se puede concluir que el masaje es más cotizado por los hombres que por las mujeres.

También se puede concluir que el Hidro es el servicio que más vende entre las 3 sucursales.

Ejercicio 2

A continuación, se muestra el flujo de datos.



En los agregados se selecciona todas las columnas exceptuando el “id” y se selecciona la opción de suma

El resultado se almacena en diferentes hojas de un mismo archivo Excel, los resultados son:

San Miguel:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rosas	Claveles	Macetas	Tierra	Girasoles	Hortensia	Globos	Tarjetas	fOrquÃ-di	CarmesÃ-	Lirios	Aurora	Tulipanes	ListÃ³n
2	157	137	141	141	150	157	151	143	158	158	160	160	149	149

San Salvador:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rosas	Claveles	Macetas	Tierra	Girasoles	Hortensia	Globos	Tarjetas	fOrquÃ-di	CarmesÃ-	Lirios	Aurora	Tulipanes	ListÃ³n
2	612	350	392	368	371	374	587	384	380	353	365	384	357	690

Santa Ana:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rosas	Claveles	Macetas	Tierra	Girasoles	Hortensia	Globos	Tarjetas	fOrquÃ-di	CarmesÃ-	Lirios	Aurora	Tulipanes	ListÃ³n
2	176	246	245	236	266	243	154	252	259	236	270	260	247	136

El Salvador (unión de los 3 departamentos anteriores):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Rosas	Claveles	Macetas	Tierra	Girasoles	Hortensia	Globos	Tarjetas	fOrquÃ-di	CarmesÃ-	Lirios	Aurora	Tulipanes	ListÃ³n
2	945	733	778	745	787	774	892	779	797	747	795	804	753	975

Porcentaje de realización: 100%

Análisis

San Miguel:

La cantidad de unidades vendidas por tipo de producto se mantienen bastante cercanas las unas de la otras, sin ver un producto que sobresalga demasiado por encima de los demás.
La variación en las ventas oscila entre las 137 (claveles) y 160 (Lirio, Aurora) unidades.
Dada la información anterior se recomienda crear diversas combinaciones, tratando de que sean llamativas a los clientes, no se recomienda forzar la inclusión de los clavales en dichas combinaciones, ya que a pesar de que es el producto menos vendido, la diferencia con el producto más vendido no es demasiada significativa.

San Salvador:

A diferencia de San Miguel, en San Salvador la diferencia de ventas entre el producto mas y menos vendido es más significativa.
La variación oscila entre 690 (Listón) y 353 (Carmesí) unidades vendidas.
Debido a lo anterior se recomienda crear combinaciones en las que se unan algunos de los productos mas vendidos con los menos vendidos, y se evite combinar los productos menos vendidos entre sí. Por ejemplo, se podría crear un arreglo con Carmesí, Rosas y Listón.

Santa Ana:

En Santa Ana tenemos un caso intermediario entre San Miguel y San Salvador, ya que la diferencia de ventas de los productos es significativa, pero no al nivel de San Salvador.
Entre 136 (Listón) y 270 (Lirios) unidades varían las ventas.
Las recomendaciones son crear diversas combinaciones juntando los productos más y menos vendidos, al igual que en San Salvador.

El Salvador:

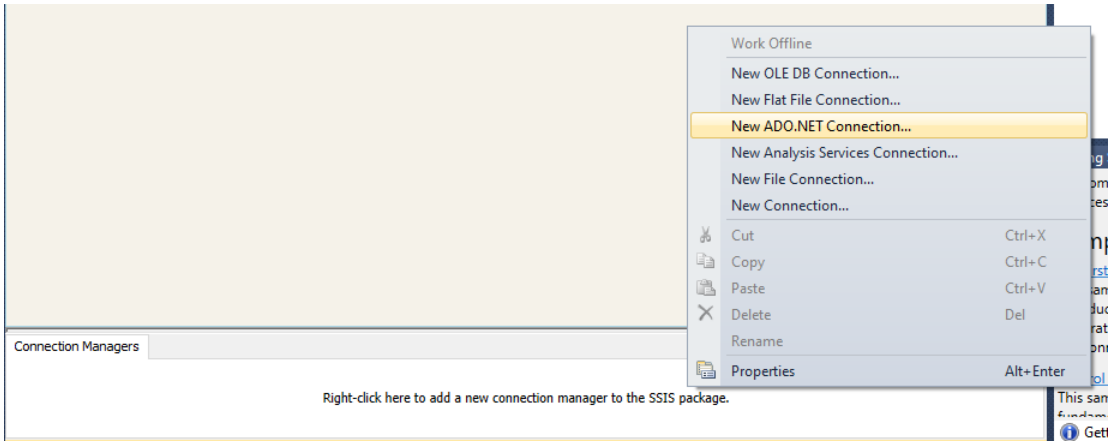
Para esto se ha tomado como una aproximación a nivel nacional, la suma de los 3 departamentos anteriormente mencionados. Con esto se obtiene y caso bastante parecido al de Santa Ana.
Las ventas oscilan entre 733 (Claveles) y 975 (Listón) unidades.
Se recomienda la creación de combinaciones variadas tratando de mezclar los productos más y menos vendidos, para así, impulsar las ventas de los productos menos vendidos.

Ejercicio 3

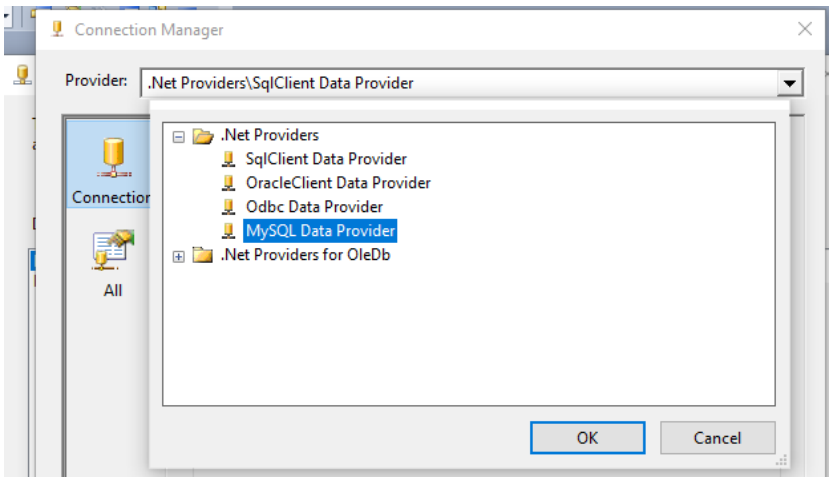
Porcentaje de realización: 100%

Para realizar el ejercicio, se añaden las conexiones pertinentes tanto a "*MySQL"* como a "*SQL Server"*, el proceso es similar, y para fines de demostración de la realización de este desafío se muestra cómo se realiza la conexión a *"MySQL"*:

Se añade una conexión de ADO.NET:



Para poder conectar con *"MySQL"* es necesario instalar el proveedor .NET indicado, es el siguiente, de esta manera podremos conectar a una base de datos:



La configuración se realiza en el apartado All se puede observar en la imagen anterior debajo de *"Connection"* y la tabla de propiedades y valores se configuran de la siguiente manera:

Property	Value
database	db_roaming_costarica
Managed Provider	MySql.Data.MySqlClient
port	3306
server	localhost
user id	root

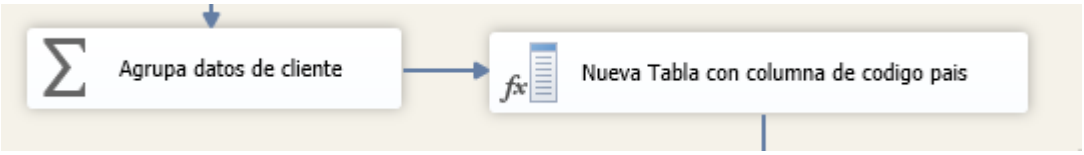
Nota: Para extraer los datos se debe realizar una consulta a la tabla requerida, esto debido a posibles incompatibilidades.

Y de esta manera conectamos satisfactoriamente a la base de datos en ***"MySQL"***, ahora agregaremos la conexión a la base de datos en SQL SERVER pasos que ya se han visto en clase por lo tanto no los añadiremos como procedimiento, las conexiones quedan de la siguiente manera:

Posteriormente juntamos los datos de ambas bases de datos, para los datos ya unidos ahora juntamos las facturaciones acumuladas para un cliente (sin repetir cliente se suman las facturaciones):

numero_telefono	numero_telefono	Group by
sexo	sexo	Group by
estado	estado	Group by
monto_facturacion	monto_facturacion	Sum
nombre_tipo	nombre_tipo	Group by

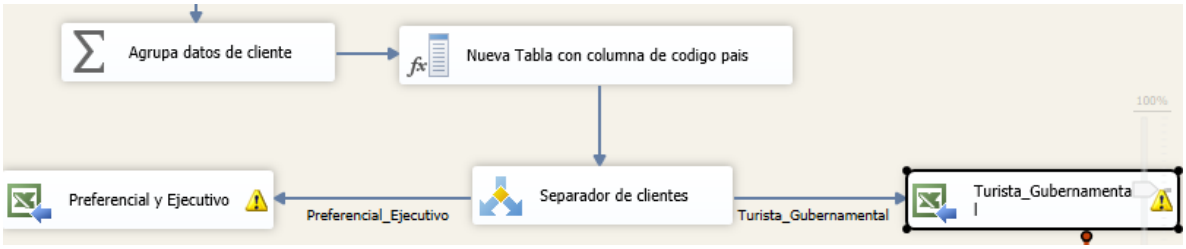
Posteriormente para cada dato ya unificado, se formatea cada dato y además se añade la nueva columna "Codigo_Pais":



Derived Column Name	Derived Column	Expression
nombre_completo	<add as new column>	nombres + apellidos
codigo_pais	<add as new column>	UPPER(SUBSTRING(codigo_cliente,1,2))
codigo_cliente	Replace 'codigo_cliente'	UPPER(codigo_cliente)
dui	Replace 'dui'	SUBSTRING(dui,1,8) + "-" + SUBSTRING(dui,9,1)
nit	Replace 'nit'	SUBSTRING(nit,1,4) + "-" + SUBSTRING(nit,5,6) + "-" + S...
sexo	Replace 'sexo'	sexo == "M" ? "MASCULINO" : "FEMENINO"
numero_telefono	Replace 'numero_telefono'	SUBSTRING(numero_telefono,1,4) + "-" + SUBSTRING(n...
estado	Replace 'estado'	SUBSTRING(numero_telefono,1,4) + "-" + SUBSTRING(n...

Se puede observar la columna ***"Código país"*** en la que se extrae los primeros 2 caracteres de del código cliente.

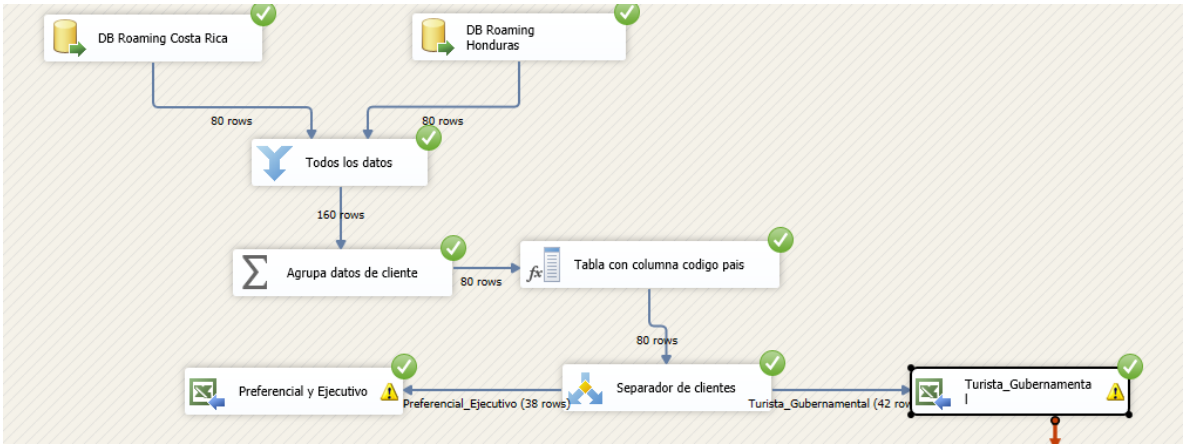
Y por último solo se añade un separador condicional, indicando que tipos de datos irán para cada archivo:



Las condiciones son las siguientes:

Order	Output Name	Condition
1	Preferencial_Ejecutivo	idtipocliente == 1 || idtipocliente == 2
2	Turista_Gubernamental	idtipocliente == 3 || idtipocliente == 4

Resultado



Excel - Cliente Preferencial y Ejecutivo

	A	B	C	D	E	F	G	H	I	J	K
1	CODIGO CLIENTE	DUI	NIT	NOMBRE COMPLETO	SEXO	TELEFONO	CODIGO PAIS	ESTADO	MONTO FACTURACION	TIPO CLIENTE	
2	SV-MM47693	00310459-4	0702-240253-	MANUEL ANTONIOMEJIA	FEMENINO	6235-7596	SV	INACTIVO	89.80	CLIENTE PREFERENCIAL	
3	SV-HE47712	01771937-7	0614-020980-	ERICK STEFFANHERNANDEZ	FEMENINO	7695-5228	SV	ACTIVO	364.60	CLIENTE EJECUTIVO	
4	SV-RJ47679	00088387-2	1416-271278-	JUAN DAVIDREYES SALAZAR	FEMENINO	7071-0453	SV	INACTIVO	292.40	CLIENTE EJECUTIVO	
5	SV-DF22100	02377745-6	0614-180648-	FRANCISCO ENRIQUEDIAZ R	FEMENINO	6523-6600	SV	ACTIVO	110.40	CLIENTE EJECUTIVO	
6	SV-CH47710	01695478-8	1324-220868-	HECTOR MANUELCHACON A	FEMENINO	7778-1096	SV	ACTIVO	82.60	CLIENTE EJECUTIVO	
7	SV-MS47711	04326581-9	1123-011290-	SAMUEL GERARDOMEDRAN	FEMENINO	7869-2068	SV	ACTIVO	219.40	CLIENTE EJECUTIVO	
8	SV-PG47677	00927645-4	0522-200472-	GUILLERMOPOCASANGRE H	FEMENINO	7789-4991	SV	ACTIVO	272.80	CLIENTE EJECUTIVO	
9	SV-ME47627	02454429-6	0607-270265-	EVELYN DE LOS ANGELESMA	FEMENINO	7070-4632	SV	ACTIVO	395.80	CLIENTE EJECUTIVO	
10	SV-LR47673	02054051-7	0614-190981-	RICARDO ERNESTOLARA CA	FEMENINO	7985-7656	SV	INACTIVO	123.40	CLIENTE PREFERENCIAL	
11	SV-GC42266	01564617-0	0511-121181-	CARLOS RICARDOGHIRINGH	FEMENINO	7247-1706	SV	INACTIVO	314.20	CLIENTE PREFERENCIAL	
12	SV-RR5909	03365212-9	0610-150982-	ROBERTORAMIREZ DEODAN	FEMENINO	7536-9733	SV	ACTIVO	43.60	CLIENTE PREFERENCIAL	
13	SV-PR20278	00374800-4	0821-220252-	RAUL ALBERTOPINEDA DIAZ	FEMENINO	6278-7729	SV	ACTIVO	302.80	CLIENTE EJECUTIVO	
14	SV-QA47695	00619610-0	0614-290358-	ANA DORA ALICIAQUINTAN	FEMENINO	6271-6439	SV	ACTIVO	134.20	CLIENTE PREFERENCIAL	
15	SV-BJ15868	01167192-9	1217-210658-	JOSE ATILIOBENITEZ PARAD	FEMENINO	6207-4477	SV	ACTIVO	308.60	CLIENTE EJECUTIVO	
16	SV-CB47683	01062229-4	1109-250583-	BLANCA MARIBELCRUZ AMA	FEMENINO	7886-6026	SV	ACTIVO	335.40	CLIENTE EJECUTIVO	
17	SV-PA13188	00466902-6	1011-280273-	ANGEL VICTORPINO MERIN	FEMENINO	7923-2040	SV	ACTIVO	156.80	CLIENTE PREFERENCIAL	
18	SV-FR47663	01075769-1	0614-240560-	ROBERTO ENRIQUEFONG HE	FEMENINO	6282-3380	SV	INACTIVO	561.80	CLIENTE PREFERENCIAL	
19	SV-CR47598	02593311-3	0614-200854-	ROSA MARGARITACRUZ GU	FEMENINO	6243-7362	SV	ACTIVO	363.40	CLIENTE EJECUTIVO	
20	SV-LC6418	01421695-6	0614-031062-	CESAR MAURICIOLOPEZ GU	FEMENINO	6305-5568	SV	ACTIVO	86.60	CLIENTE EJECUTIVO	
21	SV-SM1054	02408291-9	0203-051068-	MARITZA CAROLINASALAZA	FEMENINO	6447-8435	SV	ACTIVO	1133.60	CLIENTE PREFERENCIAL	
22	SV-PJ19861	02726756-8	0210-170964-	JOSE MAURICIOPINEDA MU	FEMENINO	6232-1846	SV	ACTIVO	369.80	CLIENTE EJECUTIVO	

Excel - Cliente Turista y Gubernamental

	A	B	C	D	E	F	G	H	I	J	K
1	CODIGO CLIENTE	DUI	NIT	NOMBRE COMPLETO	SEXO	TELEFONO	CODIGO PAIS	ESTADO	MONTO FACTURACION	TIPO CLIENTE	
2	SV-SR47678	00542374-0	0210-150150	RICARDO ANTONIOSILIEZAR SA	FEMENINO	6293-0678	SV	ACTIVO	61.20	CLIENTE GUBERNAMENTAL	
3	SV-ME47714	00449568-9	1009-100279	ERIC ALEXANDERMEJIA MOLIN	FEMENINO	7083-1594	SV	ACTIVO	61.40	CLIENTE GUBERNAMENTAL	
4	SV-FF47704	01054713-5	0511-070282	FRANCISCO STEVEFLORES SARI	FEMENINO	7921-1113	SV	ACTIVO	70.80	CLIENTE TURISTA	
5	SV-VS5897	00141561-4	0610-091159	SALVADORVASQUEZ MELARA	FEMENINO	6333-5201	SV	ACTIVO	334.40	CLIENTE GUBERNAMENTAL	
6	SV-LE47654	01945136-5	0614-181071	EVELYN ELIZABETHLOPEZ DE CA	FEMENINO	7833-2389	SV	ACTIVO	183.40	CLIENTE TURISTA	
7	SV-AL47672	00492838-9	0210-020382	LUIS ALONSOALVAREZ HERNAI	FEMENINO	7663-1438	SV	INACTIVO	301.80	CLIENTE TURISTA	
8	SV-EM47658	00374984-9	0619-030179	MARVIN VITELIOERAZO VASQU	FEMENINO	7786-5956	SV	INACTIVO	93.40	CLIENTE TURISTA	
9	SV-MF47700	00889131-6	0614-091258	FRANCISCO REMBERTOMIXCO	FEMENINO	6276-3473	SV	INACTIVO	155.20	CLIENTE TURISTA	
10	SV-CC47631	02216998-5	0112-260574	CARLOS ADALBERTOCASTRO A	FEMENINO	6541-4440	SV	ACTIVO	85.60	CLIENTE GUBERNAMENTAL	
11	SV-CO47689	02473300-7	1010-290665	OSCAR MAURICIOCARRILLO TU	FEMENINO	6288-8115	SV	ACTIVO	127.80	CLIENTE TURISTA	
12	SV-RC47664	00569371-9	0614-020383	CARLOS RODOLFOROSALES MA	FEMENINO	6566-6345	SV	ACTIVO	367.60	CLIENTE TURISTA	
13	SV-GM47660	01129927-2	0805-120572	MANUEL ENRIQUEGRANDE CA	FEMENINO	6284-5726	SV	INACTIVO	186.20	CLIENTE GUBERNAMENTAL	
14	SV-PC47670	01040243-2	0714-230573	CRUZ MARINAPALACIOS DE GU	FEMENINO	6203-9568	SV	ACTIVO	194.60	CLIENTE TURISTA	
15	SV-HG47715	03816206-4	0614-241087	GUSTAVO RAFAELHERNANDEZ	FEMENINO	7104-8822	SV	ACTIVO	363.00	CLIENTE TURISTA	
16	SV-DR47657	00864458-1	0312-050365	ROSSANA ELISABETHDIAZ DE C	FEMENINO	7457-5307	SV	ACTIVO	283.80	CLIENTE TURISTA	
17	SV-CM1664	01596553-2	0210-091163	MARTA ALICIACABRERA MART	FEMENINO	6280-5523	SV	ACTIVO	202.80	CLIENTE GUBERNAMENTAL	
18	SV-RM47671	00738151-2	0614-310570	MANUEL ALFREDORIVERA MEI	FEMENINO	7189-5007	SV	ACTIVO	302.80	CLIENTE TURISTA	
19	SV-HM47684	04432219-3	0614-120391	MILTON SAMUELHERNANDEZ (	FEMENINO	7581-2813	SV	INACTIVO	192.40	CLIENTE GUBERNAMENTAL	
20	SV-BA47686	00119483-2	0614-251259	ANA IRMABENITEZ ARGUETA	FEMENINO	6270-5458	SV	ACTIVO	390.80	CLIENTE TURISTA	
21	SV-NS47705	00115399-5	0604-090668	SIMEONNAVARRETE LEONOR	FEMENINO	6124-2645	SV	ACTIVO	199.00	CLIENTE TURISTA	
22	SV-LJ47662	03975969-4	0612-240888	JORGE MATEOLOPEZ URQUILLA	FEMENINO	7832-6391	SV	ACTIVO	205.80	CLIENTE TURISTA	