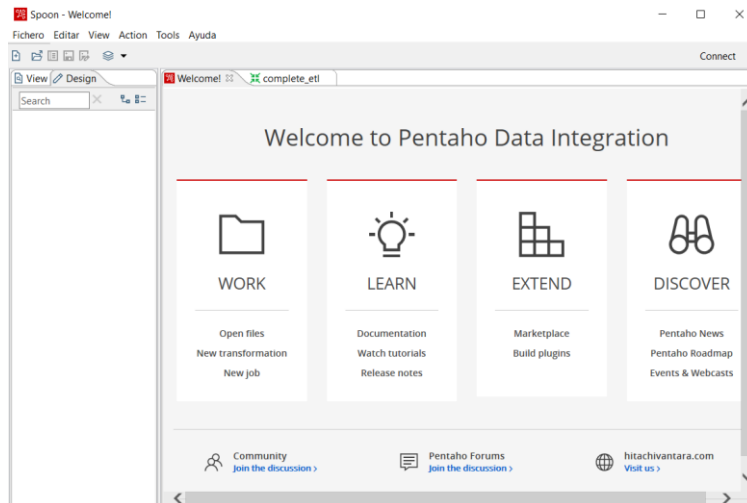


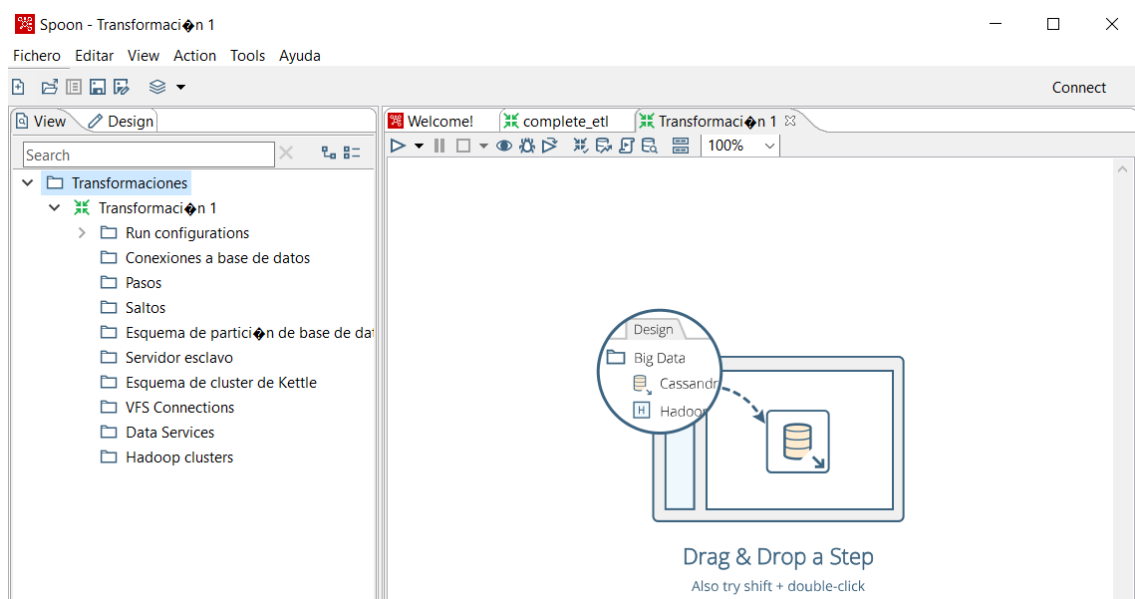
Task3. Data transformation

In this task I am going to show how I have done the ETL process for the Airlines database using Pentaho data integration.

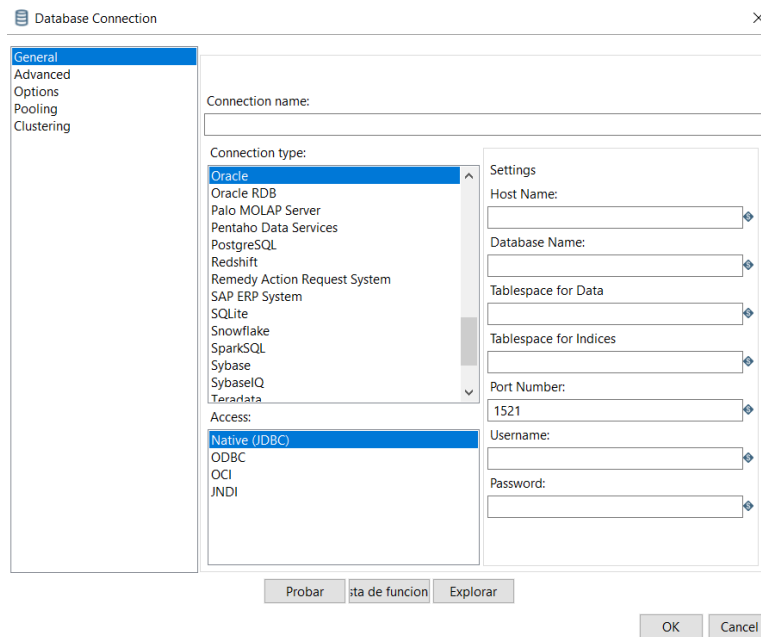
First, we have to open the data integration program.



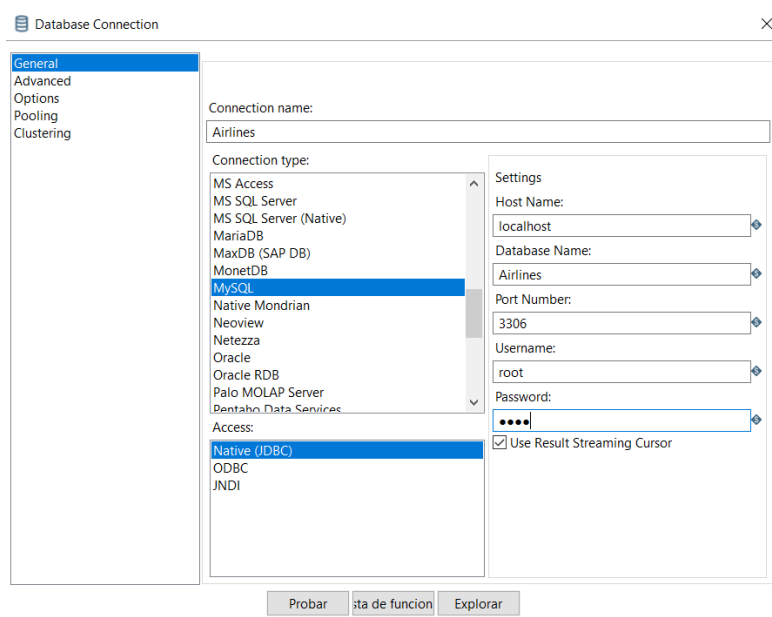
Now I am going to create a new transformation to start the ETL process.



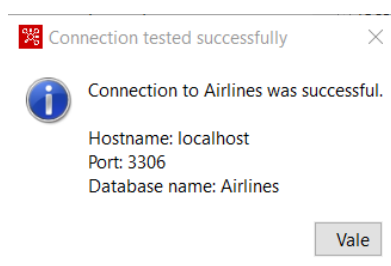
The next step is to connect it to the Airlines database. I am going to do that by right clicking on 'Conexiones a bases de datos' -> New.



After clicking it appears this window. In it we have to put the database login info as we did in the schema-workbench.



When the info is on the connection we click on 'Probar' to check if the connection can be made. To see that everything is ok we have to see this window.



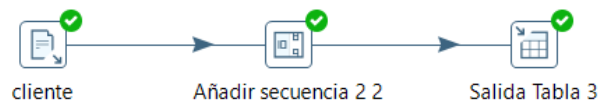
Now that the connection can be made we have to click in 'Ok' and we can start the ETL process.

-ETL

The first thing we are going to do is load the dimension tables.

We are going to start by uploading the client table.

-Customer



In the first step we have the entry of the CSV table.

The screenshot shows the 'CSV file input' dialog box. The 'Step name' is 'cliente'. The 'Filename' is 'D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D'. The 'Delimiter' is ';'. The 'Enclosure' is '"'. The 'NIO buffer size' is '50000'. The 'Lazy conversion?' checkbox is unchecked. The 'Header row present?' checkbox is checked. The 'Add filename to result' checkbox is unchecked. The 'The row number field name (optional)' field is empty. The 'Running in parallel?' checkbox is unchecked. The 'New line possible in fields?' checkbox is unchecked. The 'Format' is 'mixed'. The 'File encoding' is 'UTF-8'. Below the dialog box, there is a table with 11 columns: #, Name, Type, Format, Length, Precision, Currency, Decimal, Group, Trim type, and a blank column. The table contains 11 rows of data for the 'cliente' table.

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type	
1	customer_key	Integer	#	15	0	€	,	-	ninguno	
2	customer_name	String		9		€	,	-	ninguno	
3	customer_address	String		15		€	,	-	ninguno	
4	customer_city	String		17		€	,	-	ninguno	
5	customer_state	String		14		€	,	-	ninguno	
6	customer_zip	Integer	#	15	0	€	,	-	ninguno	
7	customer_type	String		13		€	,	-	ninguno	
8	customer_income	String		8		€	,	-	ninguno	
9	customer_birth_date	String		19		€	,	-	ninguno	
10	customer_marital	String		10		€	,	-	ninguno	
11	customer_sex	String		1		€	,	-	ninguno	

In this step we have to select the address of the CSV file we have for the client, we also have to put the semicolon separator because it is a CSV file what we are importing. After that we click on "Traer campos" and then in "Vale".

In the next step we added a add sequence block because we want to create an autogenerated value in order to create 'idcliente' so we can identify the clients in the fact table.

Obtener valor de la secuencia de la base de datos

Nombre de paso: Añadir secuencia 2 2

Nombre de valor: idcliente

Utilizar una base de datos para generar la secuencia

¿Utilizar base de datos para obtener secuencia? ☐

Conexión: Airlines [Editar...] [Nuevo...] [Wizard...]

Nombre de esquema: [Schemas...]

Nombre de secuencia: SEQ_ [Sequences...]

Utilizar un contador de la transformación para generar la secuencia

¿Utilizar contador para calcular secuencia? ☒

Nombre contador (opcional):

Valor inicial: 1

Incremento: 1

Valor máximo: 999999999

[Help] [Vale] [Cancelar]

Once we have the 'idcliente' attribute ready' the las step is to upload the table to the cliente table in MySql server.

Salida de Tabla

Nombre paso: Salida Tabla 3

Conexión: Airlines [Editar...] [Nuevo...] [Wizard...]

Esquema destino: Airlines [Examinar...]

Tabla destino: cliente [Examinar...]

Tamaño de transacción (commit): 1000

Vaciar tabla ☒

Ignorar errores de inserción ☐

Specify database fields ☒

Main options Database fields

Fields to insert:

#	Table field	Stream field
1	customer_key	customer_key
2	customer_name	customer_name
3	customer_address	customer_address
4	customer_city	customer_city
5	customer_state	customer_state
6	customer_zip	customer_zip
7	customer_type	customer_type
8	customer_income	customer_income
9	customer_birth_date	customer_birth_date
1..	customer_marital	customer_marital
1..	customer_sex	customer_sex
1..	idcliente	idcliente

[Get fields] [Enter field mapping]

[Help] [Vale] [Cancelar] [SQL]

To do that we have to select the destiny table in the field "Tabla destino" and then set the fields to insert. We can see that the "idcliente" field has been inserted correctly.

If we Run the program, we can do a SQL query in the Airlines database to check that the data has been inserted.

Limit to 2000 rows

```

1 • select * from Airlines.canal;
2 • select * from Airlines.aeropuerto;
3 • select * from Airlines.cliente;
4 • select * from Airlines.tarifa;
5 • select * from Airlines.vuelo;
6 • select * from Airlines.trayecto order by flown_key,cliente;
7 • select * from Football.player;

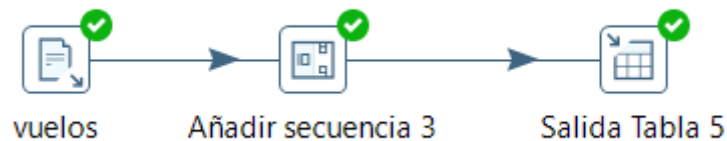
```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content: |

	idcliente	customer_key	customer_name	customer_address	customer_city	customer_state	customer_zip	customer_type	customer_inco
1	1	Anderson	1607 Shady Lane	Birmingham	Alabama	40928	Employed	\$102.000	
2	2	Antoni	3859 Shady Lane	Tuscaloosa	Alabama	35294	Employed	\$35.000	
3	3	Appleby	1923 Shady Lane	Anchorage	Alaska	58358	Employed	\$47.000	
4	4	Ashby	9369 Shady Lane	Juneau	Alaska	90421	Employed	\$94.000	
5	5	Barr	7593 Shady Lane	Flagstaff	Arizona	67536	Employed	\$93.000	
6	6	Barrett	5332 Shady Lane	Phoenix	Arizona	88392	Employed	\$117.000	
7	7	Bennett	4116 Shady Lane	Little Rock	Arkansas	23848	Self Employed	\$35.000	
8	8	Boone	3100 Shady Lane	Midville	Arkansas	88536	Employed	\$93.000	
9	9	Clarke	7808 Shady Lane	San Diego	California	39238	Employed	\$61.000	
10	10	Clewett	3997 Shady Lane	Red Bluff	California	37374	Employed	\$87.000	
11	11	Cluster	8640 Shady Lane	Denver	Colorado	61892	Military	\$44.000	
12	12	Coghlin	7143 Shady Lane	Steamboat S...	Colorado	92682	Self Employed	\$44.000	
13	13	Davis	8765 Shady Lane	Hartford	Connecticut	94452	Employed	\$91.000	

-Flights

Now I am going to add the flight table, we are going to do the same as in the customers one.



In the first step we have the block to import the CSV file.

CSV file input

Step name: vuelos

Filename: D:\Documentos\Master\temariomaster\p4\DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☐

Header row present? ☒

Add filename to result ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	flight_key	Integer	#	15	0	€	.	.	ninguno
2	sched_depart	String		5		€	.	.	ninguno
3	sched_arrival	String		5		€	.	.	ninguno
4	airplane_type	String		8		€	.	.	ninguno
5	seat_capacity	Integer	#	15	0	€	.	.	ninguno
6	first_class_capacity	Integer	#	15	0	€	.	.	ninguno
7	business_capacity	Integer	#	15	0	€	.	.	ninguno
8	coach_capacity	Integer	#	15	0	€	.	.	ninguno

Help | Vale | Traer Campos | Previsualizar | Cancelar

After that I am going to generate the 'idvuelo' attribute the same way I did in the client table.

Obtener valor de la secuencia de la base de datos

Nombre de paso: Añadir secuencia 3

Nombre de valor: idvuelo

Utilizar una base de datos para generar la secuencia

¿Utilizar base de datos para obtener secuencia? ☐

Conexión: Airlines

Nombre de esquema:

Nombre de secuencia: SEQ_

Utilizar un contador de la transformación para generar la secuencia

¿Utilizar contador para calcular secuencia? ☒

Nombre contador (opcional):

Valor inicial: 1

Incremento: 1

Valor máximo: 99999999

Help Vale Cancelar

And in the last step I am going to choose the destiny table in the Airlines database with the 'idvuelo' generated.

Salida de Tabla

Nombre paso: Salida Tabla 5

Conexión: Airlines

Esquema destino: Airlines

Tabla destino: vuelo

Tamaño de transacción (commit): 1000

Vaciar tabla: ☒

Ignorar errores de inserción: ☐

Specify database fields: ☒

Main options Database fields

Fields to insert:

Table field	Stream field
1 flight_key	flight_key
2 sched_dep...	sched_depart
3 sched_arriv...	sched_arrival
4 airplane_ty...	airplane_type
5 seat_capaci...	seat_capacity
6 first_class_c...	first_class_ca...
7 business_ca...	business_cap...
8 coach_capa...	coach_capaci...
9 idvuelo	idvuelo

Get fields Enter field mapping

Help Vale Cancelar SQL

To check that it has worked we are going to do the SQL query.

Query 1 plays SQL File 4* x

Limit to 2000 rows

```
1 • select * from Airlines.canal;
2 • select * from Airlines.aeropuerto;
3 • select * from Airlines.cliente;
4 • select * from Airlines.tarifa;
5 • select * from Airlines.vuelo;
6 • select * from Airlines.trayecto order by flown_key,cliente;
7 • select * from Football.player;
```

Result Grid

Filter Rows:

Edit: Export/Import: Wrap Cell Content: FX

	idvuelo	flight_key	sched_depart	sched_arrival	airplane_type	seat_capacity	first_class_capacity	business_capacity	coach_capacity
31	31	10:35	12:05	Super 80	150	14	0	136	
32	32	10:40	12:21	727	130	16	0	114	
33	33	10:45	12:17	727	130	16	0	114	
34	34	10:50	12:07	DC-10	300	28	28	244	
35	35	10:55	13:30	DC-10	300	28	28	244	
36	36	11:00	13:51	727	130	16	0	114	
37	37	11:05	12:21	727	130	16	0	114	
38	38	11:10	14:00	DC-10	300	28	28	244	
39	39	11:15	14:03	DC-10	300	28	28	244	
40	40	11:20	12:10	DC-10	300	28	28	244	
41	41	11:25	14:02	727	130	16	0	114	
42	42	11:30	13:12	DC-10	300	28	28	244	
43	43	11:35	13:45	727	130	16	0	114	
44	44	11:40	12:35	Super 80	150	14	0	136	

Result Grid

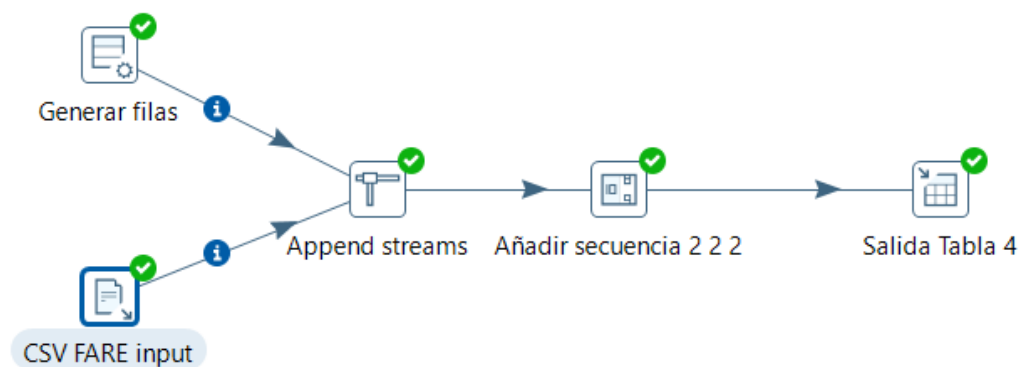
Form Editor

Field Types

Query Stats

-Fare

Now we are going to import the fare table. With this table we have a problem because the fare CSV we have four kinds of fare_type and in the frequentflyer table we can see that this number goes up to 20 kind of fares. This could mean that the data we have doesn't have all the fare types or that this kind of fares have been eliminated. It could also mean that the people that took the data didn't do it properly so for this kind of data whose fare_key goes up of 4 we are going to create a row in the fare table with a 5 idfare and a -1 in fare_key to store there the incorrect data, this way we don't lose it.



The CSV fare input is the same block as in the cases explained before.

CSV file input

Step name: CSV FARE input

Filename: D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☐

Header row present? ☒

Add filename to result? ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	fare_class_key	Integer	#	15	0	€	,	.	ninguno
2	fare_class_code	String		1		€	,	.	ninguno
3	fare_class_description	String		8		€	,	.	ninguno
4	restriction_type	String		14		€	,	.	ninguno

Help Vale Traer Campos Previsualizar Cancelar

Now I have the generate row block.

Generar filas

Nombre paso:

Límite:

Never stop generating rows: ☐

Interval in ms (delay):

Current row time field name:

Previous row time field name:

Campos:

#	Nombre	Tipo	Formato	Longitud	Precisión	Moneda	Decimal	Grupo	Valor	Set empty
1	fare_class_key	Integer	#	15	0	€	,	.	-1	N
2	fare_class_code	String		1		€	,	.		N
3	fare_class_description	String		8		€	,	.		N
4	restriction_type	String		14		€	,	.		N

[Help](#) [Vale](#) [Previsualizar](#) [Cancelar](#)

This row creates a row where the value of 'fare_class_key' is -1.

After that we find the append streams block, it unifies the CSV and the generated row from the block before.

Append

Nombre de paso:

Head hop:

Tail hop:

[Help](#) [Vale](#) [Cancelar](#)

After that we have the generate sequence once again.

Obtener valor de la secuencia de la base de datos

Nombre de paso:

Nombre de valor:

Utilizar una base de datos para generar la secuencia

¿Utilizar base de datos para obtener secuencia? ☐

Conexión: [Editar...](#) [Nuevo...](#) [Wizard...](#)

Nombre de esquema: [Schemas...](#)

Nombre de secuencia: [Sequences...](#)

Utilizar un contador de la transformación para generar la secuencia

¿Utilizar contador para calcular secuencia? ☒

Nombre contador (opcional):

Valor inicial:

Incremento:

Valor máximo:

[Help](#) [Vale](#) [Cancelar](#)

And in the last place the output for the Airlines database.

Salida de Tabla

Nombre paso: Salida Tabla 4

Conexión: Airlines

Esquema destino: Airlines

Tabla destino: tarifa

Tamaño de transacción (commit): 1000

Vaciar tabla: ☒

Ignorar errores de inserción: ☐

Specify database fields: ☒

Main options Database fields

Fields to insert:

#	Table field	Stream field
1	fare_class_k...	fare_class_key
2	fare_class_c...	fare_class_co...
3	fare_class_...	fare_class_de...
4	restriction_...	restriction_ty...
5	idtarifa	idtarifa

Get fields

Enter field mapping

Help Vale Cancelar SQL

To check again we are going to do the SWL query.

Query 1 plays SQL File 4*

Limit to 2000 rows

```

1 • select * from Airlines.canal;
2 • select * from Airlines.aeropuerto;
3 • select * from Airlines.cliente;
4 • select * from Airlines.tarifa;
5 • select * from Airlines.vuelo;
6 • select * from Airlines.trayecto order by flown_key,cliente;
7 • select * from Football.player;

```

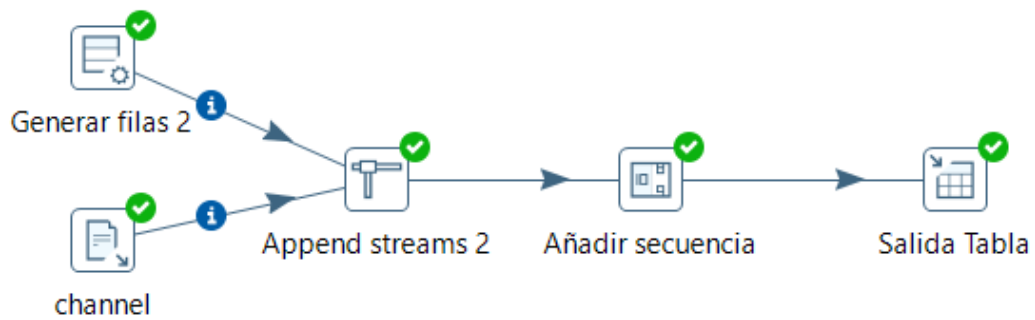
Result Grid

	idtarifa	fare_class_key	fare_class_code	fare_class_description	restriction_type
1	1	Y	Economy	None	
2	2	J	Business	None	
3	3	F	First	None	
4	4	X	Discount	30 Day Advance	
5	-1	NULL	NULL	NULL	
*	NULL	NULL	NULL	NULL	

We can see how the row has been created.

-Channel

Now we are going to import the channel CSV, in this case it happens the same thing as before so to fix it we are also going to generate a row.



First we import the CSV.

CSV file input

Step name: channel

Filename: D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☐

Header row present? ☒

Add filename to result? ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	channel_key	Integer	#	15	0	€	,	.	ninguno
2	channel_name	String		11		€	,	.	ninguno

Buttons: Help, Vale, Traer Campos, Previsualizar, Cancelar

After that we generate the row with value -1 for the 'channel_key' attribute.

Generar filas

Nombre paso: Generar filas 2

Límite: 1

Never stop generating rows? ☐

Interval in ms (delay): 5000

Current row time field name: now

Previous row time field name: FiveSecondsAgo

Campos:

#	Nombre	Tipo	Formato	Longitud	Precisión	Moneda	Decimal	Grupo	Valor	Set empty string?
1	channel_key	Integer	#	15	0	€	,	.	-1	N
2	channel_name	String		11		€	,	.		N

Buttons: Help, Vale, Previsualizar, Cancelar

We append the data.

Append

Nombre de paso: Append streams 2

Head hop: channel

Tail hop: Generar filas 2

Buttons: Help, Vale, Cancelar

Now we add the sequence and generate the output.

Obtener valor de la secuencia de la base de datos

Nombre de paso: Añadir secuencia

Nombre de valor: idcanal_sec

Utilizar una base de datos para generar la secuencia

¿Utilizar base de datos para obtener secuencia? ☐

Conexión: Airlines

Nombre de esquema:

Nombre de secuencia: SEQ_

Utilizar un contador de la transformación para generar la secuencia

¿Utilizar contador para calcular secuencia? ☒

Nombre contador (opcional):

Valor inicial: 1

Incremento: 1

Valor máximo: 999999999

Help Vale Cancelar

Salida de Tabla

Nombre paso: Salida Tabla

Conexión: Airlines

Esquema destino: Airlines

Tabla destino: canal

Tamaño de transacción (commit): 1000

Vaciar tabla: ☒

Ignorar errores de inserción: ☐

Specify database fields: ☒

Main options Database fields

Fields to insert:

#	Table field	Stream field
1	idcanal	idcanal_sec
2	channel_name	channel_name
3	channel_key	channel_key

Get fields

Enter field mapping

Help Vale Cancelar SQL

Last we are going to check MySQL again.

Query 1 plays SQL File 4*

Limit to 2000 rows

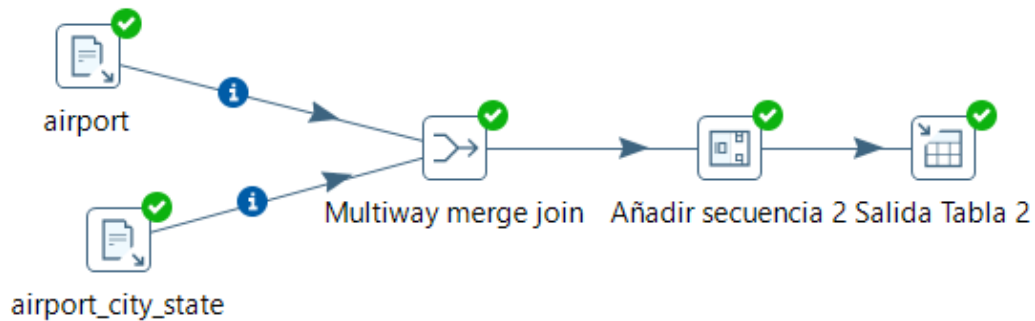
```
1 • select * from Airlines.canal;
2 • select * from Airlines.aeropuerto;
3 • select * from Airlines.cliente;
4 • select * from Airlines.tarifa;
5 • select * from Airlines.vuelo;
6 • select * from Airlines.trayecto order by flown_key,cliente;
7 • select * from Football.player;
```

Result Grid

	idcanal	channel_key	channel_name
▶	1	1	Cash
	2	2	Credit Card
	3	3	Debit Card
	4	4	PayPal
	5	-1	NULL
*	NULL	NULL	NULL

-Airport

In this table we have to unify the airport table and the airport city_state_table. To do that we have to import both CSV files and do a multiway merge join.



First we import de CSVs

CSV file input

Step name: airport_city_state

Filename: D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☒

Header row present? ☒

Add filename to result? ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	city	String		14		€	,	.	ninguno
2	state	String		2		€	,	.	ninguno

Help Vale Traer Campos Previsualizar Cancelar

CSV file input

Step name: airport

Filename: D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☒

Header row present? ☒

Add filename to result? ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	leg_key	Integer	#	15	0	€	,	.	ninguno
2	leg_name	String		15		€	,	.	ninguno
3	leg_city	String		14		€	,	.	ninguno
4	leg_type	String		13		€	,	.	ninguno
5	leg_radar_type	String		3		€	,	.	ninguno

Help Vale Traer Campos Previsualizar Cancelar

After that we have to choose where we want to unify the tables, in this case what we want is to unify them by city.

Multiway merge join

Step name: Multiway merge join

Input Step1: airport Join Keys: leg_city Select Keys

Input Step2: airport_city_state Join Keys: city Select Keys

Join Type: INNER

Help Vale Cancelar

After that we add the sequence and select the output.

Obtener valor de la secuencia de la base de datos

Nombre de paso: Añadir secuencia 2

Nombre de valor: idaeropuerto

Utilizar una base de datos para generar la secuencia

¿Utilizar base de datos para obtener secuencia? ☐

Conexión: Airlines Editar... Nuevo... Wizard...

Nombre de esquema: Schemas...

Nombre de secuencia: SEQ Sequences...

Utilizar un contador de la transformación para generar la secuencia

¿Utilizar contador para calcular secuencia? ☒

Nombre contador (opcional):

Valor inicial: 1

Incremento: 1

Valor máximo: 99999999

Help Vale Cancelar

Salida de Tabla

Nombre paso: Salida Tabla 2

Conexión: Airlines Editar... Nuevo... Wizard...

Esquema destino: Airlines Examinar...

Tabla destino: aeropuerto Examinar...

Tamaño de transacción (commit): 1000

Vaciar tabla: ☒

Ignorar errores de inserción: ☐

Specify database fields: ☒

Main options Database fields

Fields to insert:

#	Table field	Stream field
1	leg_key	leg_key
2	leg_name	leg_name
3	leg_city	leg_city
4	leg_type	leg_type
5	leg_radar_t...	leg_radar_type
6	city	city
7	state	state
8	idaeropuer...	idaeropuerto

Get fields Enter field mapping

Help Vale Cancelar SQL

Now we are going to check the data in Airlines database.

Query 1 plays SQL File 4*

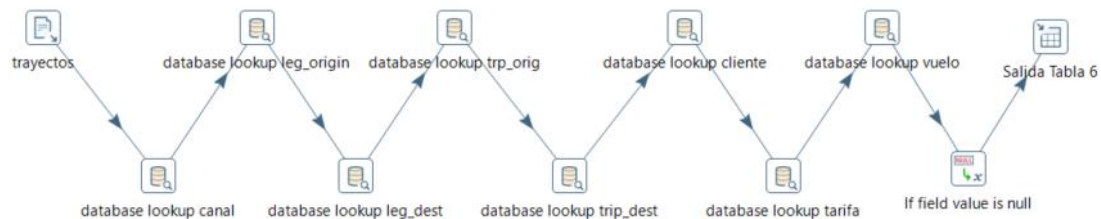
</

We can see that the merge join has been done perfectly.

Now it is time to import the fact table.

-Frequent Flyer.

To import the fact table, I have preceded following the upcoming steps.



In the first place I have imported the CSV file that has the frequent flyer information.

CSV file input

Step name: trayectos

Filename: D:\Documentos\Master\temariomaster\p4.DataAnalytics\Ejercicios\Ejercicio1.D

Delimiter: ;

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☐

Header row present? ☒

Add filename to result? ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	flown_key	Integer	#	15	0	€	,	.	ninguno
2	customer_key	Integer	#	15	0	€	,	.	ninguno
3	leg_origin_key	Integer	#	15	0	€	,	.	ninguno
4	leg_dest_key	Integer	#	15	0	€	,	.	ninguno
5	trip_origin_key	Integer	#	15	0	€	,	.	ninguno
6	trip_dest_key	Integer	#	15	0	€	,	.	ninguno
7	flight_key	Integer	#	15	0	€	,	.	ninguno
8	fare_class_key	Integer	#	15	0	€	,	.	ninguno
9	channel_key	Integer	#	15	0	€	,	.	ninguno
10	fare	Integer	#	15	0	€	,	.	ninguno
11	miles	Integer	#	15	0	€	,	.	ninguno
12	minutes_late	Integer	#	15	0	€	,	.	ninguno
13	ticket_number	Integer	#	15	0	€	,	.	ninguno

Help Vale Traer Campos Previsualizar Cancelar

Now that I have the table to work with. I have done several lookups tables to check that the key information in the fact table related to the id of the foreign key is equal to the primary key of the dimension table. If that happens it assigns the 'idvalue' from the dimension table to that value.

Let's see the client example to have a clear understanding.

Búsqueda de valor en base de datos

Nombre paso: database lookup cliente

Conexión: Airlines2

Esquema de búsqueda: Airlines

Tabla de búsqueda: cliente

Habilitar cache? ☐

Tamaño de cache en filas (0=todas): 0

Load all data from table ☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	customer_key	=	customer_key	

Valores a devolver de la tabla de búsqueda:

#	Campo	Nuevo nombre	Defecto	Tipo
1	idcliente			None

No se puede ir a la búsqueda de la...

If we pay attention to the image I am checking if the 'customer_key' from frequentlyflyer is the same as the 'customer_key' from client table. If that happens, I assign the value 'idcliente' to the output of the fact table.

I have done that for every foreign key of the fact table so I'll just show the images doing that.

Búsqueda de valor en base de datos

Nombre paso

database lookup canal

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

canal

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	channel_key	=	channel_key	

Valores a devolver de la tabla de búsqueda :

#	Campo	Nuevo nombre	Defecto	Tipo
1	idcanal			Integer

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso

database lookup leg_origin

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

aeropuerto

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	leg_key	=	leg_origin_key	

Valores a devolver de la tabla de búsqueda :

#	Campo	Nuevo nombre	Defecto	Tipo
1	idaeropuerto	idaeropuertolegor		None

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso

database lookup leg_dest

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

aeropuerto

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	leg_key	=	leg_dest_key	

Valores a devolver de la tabla de búsqueda :

#	Campo	Nuevo nombre	Defecto	Tipo
1	idaeropuerto	idaeropuertolegdest		None

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso

database lookup trp_orig

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

aeropuerto

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	leg_key	=	trip_origin_key	

Valores a devolver de la tabla de búsqueda:

#	Campo	Nuevo nombre	Defecto	Tipo
1	idaaeropuerto	idaaerpuertotrip		None

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso

database lookup trp_dest

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

aeropuerto

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	leg_key	=	trip_dest_key	

Valores a devolver de la tabla de búsqueda:

#	Campo	Nuevo nombre	Defecto	Tipo
1	idaaeropuerto	idaaerpuertotripdest		None

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso

database lookup tarifa

Conexión

Airlines2

Editar...

Nuevo...

Wizard...

Esquema de búsqueda

Airlines

Examinar...

Tabla de búsqueda

tarifa

Examinar...

Habilitar cache?

☐

Tamaño de cache en filas (0=todas)

0

Load all data from table

☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	fare_class_key	=	fare_class_key	

Valores a devolver de la tabla de búsqueda:

#	Campo	Nuevo nombre	Defecto	Tipo
1	idtarifa			Integer

No procesar la fila si la búsqueda falla

☐

Producir error si se obtienen múltiples

☐

Ordenar por

Help

Vale

Cancelar

Obtener Campos

Obtener Campos Búsqueda

Búsqueda de valor en base de datos

Nombre paso: database lookup vuelo

Conexión: Airlines2

Esquema de búsqueda: Airlines

Tabla de búsqueda: vuelo

Habilitar cache? ☐

Tamaño de cache en filas (0-todas): 0

Load all data from table ☐

La clave(s) para realizar búsqueda de valor(es):

#	Campo de tabla	Comparador	Campo1	Campo2
1	flight_key	=	flight_key	

Valores a devolver de la tabla de búsqueda:

#	Campo	Nuevo nombre	Defecto	Tipo
1	idvuelo			None

No procesar la fila si la búsqueda falla ☐

Producir error si se obtienen múltiples ☐

Ordenar por:

Help Vale Cancelar Obtener Campos Obtener Campos Búsqueda

Now that we have done all the lookups, if you remember, in the channel and fare tables we created an additional row to store the data we had in the fact table that did not match the data we have in the dimension table.

For example, when we do a lookup to that kind of value, if we are comparing and we get a fare_key of 4 in the dimension table and we get a fare_key of 19 in the fact table, to that value the pentaho data integration its going to assign a null in the key field.

That is why we use the block 'if null value'.

If field value is null

Step name: if field value is null

Replace Null for all fields

Replace by value:

Set empty string? ☐

Mask (Date):

Select fields ☒

Select value type ☐

Value types

#	Type	Replace by value	Conversion mask (Date)	Set empty string?
---	------	------------------	------------------------	-------------------

Fields

#	Field	Replace by value	Conversion mask (Date)	Set empty string?
1	idcanal	-1		N
2	idtarifa	-1		N

Help Vale Traer Campos Cancelar

With this we say, if the field 'idcanal' or 'idtarifa' has a null value assign a -1 to that field, that way we have that kind of data localized and we don't have null values.

To sum up the data integration we have to upload the processed data to the fact table in the Airlines dataset. To do that we are going to use the block we have been using before.

Salida de Tabla

Nombre paso: Salida Tabla 6

Conexión: Airlines2

Esquema destino: Airlines

Tabla destino: trayecto

Tamaño de transacción (commit): 1000

☒ Vaciar tabla

☐ Ignorar errores de inserción

☒ Specify database fields

Main options / Database fields

Fields to insert:

#	Table field	Stream field
1	flow_key	flow_key
2	cliente	idcliente
3	aeropuerto_origen_trayecto	idaeropuertolegor
4	aeropuerto_destino_trayecto	idaeropuertolegdest
5	aeropuerto_origen_itinerario	idaeropuertotripor
6	aeropuerto_destino_itinerario	idaeropuertotripdest
7	vuelo	idvuelo
8	fare	fare
9	canal	idcanal
1.	tarifa	idtarifa
1.	miles	miles
1.	minutes_late	minutes_late
1.	ticket_number	ticket_number

Get fields

Enter field mapping

Help Vale Cancelar SQL

To check that the data has been uploaded correctly we are going to do a SQL query.

Query 1 plays SQL File 4

```

1 select * from Airlines.canal;
2 select * from Airlines.aeropuerto;
3 select * from Airlines.cliente;
4 select * from Airlines.tarifa;
5 select * from Airlines.vuelo;
6 select * from Airlines.trayecto;
7 # order by flow_key, cliente;

```

Result Grid

flow_key	fare	miles	minutes_late	ticket_number	aeropuerto_origen_trayecto	aeropuerto_origen_itinerario	aeropuerto_destino_trayecto
18	157	828	0	3019	12	12	1
18	21	112	36	3019	1	12	6
24	375	1972	22	3019	6	6	12
38	303	1597	0	1063	5	5	14
38	360	1897	40	1063	14	5	11
41	99	522	0	1063	11	11	4
41	238	1255	0	1063	4	11	5
32	191	1004	0	8348	14	14	19
37	239	1257	0	8348	19	19	14
51	316	1665	0	3904	14	14	11
54	136	714	0	3904	11	11	17
54	288	1515	0	3904	17	11	14
44	332	1745	0	954	5	5	4

trayecto 60

Query 1 plays SQL File 4

```

1 select * from Airlines.canal;
2 select * from Airlines.aeropuerto;
3 select * from Airlines.cliente;
4 select * from Airlines.tarifa;
5 select * from Airlines.vuelo;
6 select * from Airlines.trayecto;
7 # order by flow_key, cliente;

```

Result Grid

rto_origen_trayecto	aeropuerto_origen_itinerario	aeropuerto_destino_trayecto	aeropuerto_destino_itinerario	cliente	canal	tarifa	vuelo
12	1	6	1	-1	-1	20	
12	6	6	1	-1	-1	94	
6	12	12	1	-1	-1	6	
5	14	11	1	-1	-1	1	
5	11	11	1	-1	-1	76	
11	4	5	1	-1	2	15	
11	5	5	1	-1	-1	84	
14	19	19	1	2	-1	11	
19	14	14	1	2	4	10	
14	11	11	1	4	-1	16	
11	17	14	1	4	4	3	
11	14	14	1	4	-1	92	
5	4	7	1	-1	2	24	

In here we can check the -1 value for the channel and fare bad values.

