



Caso 103 Quantity

1 Casos ETL Modelado de Datos.

Jose Ignacio González Gómez

Departamento de Economía, Contabilidad y Finanzas - Universidad de La Laguna

www.jggomez.eu

V.21

Ejercicio Basado: *Desarrollado por jggomez*

Enunciado

Partimos de un fichero CSV con el registro de existencias en almacén en un formato no normalizado donde podemos observar que existen un rango de datos que contiene filas en blanco, otras con valores NA, y otros errores que exigen una depuración y limpieza de esta fuente antes de incorporarla a un modelo de datos.

	A	B	C	D	E
1	Product ID	Producto	Made in	Batch code	Quantity
2	1	Producto1	China	1220C	370
3	2	Producto2	Francia	1314C	2884
4	3	Producto3	India	1996I	2248
12					
13					
29	NA	NA	NA	NA	NA
30	NA	NA	NA	NA	NA
31	30	Producto30	India	1059I	2142
32	31	Producto31	Turquia	1113P	671

Se pide

Con la información no estructurada del fichero original deseamos crear una conexión y transformarla en un formato estructurado usando Power Query obteniendo una tabla similar a la siguiente.

	A ^B Product ID	A ^B Producto	A ^B Made in	A ^B Batch code	i ² Quantity
1	1	Producto1	China	1220C	370
2	2	Producto2	Francia	1314C	2884
3	3	Producto3	India	1996I	2248
4	4	Producto4	Turquia	1348P	2878
5	5	Producto5	Reino Unido	1904U	3201
95	111	Producto111	China	1524C	3262
96	112	Producto112	Francia	1080C	797
97	113	Producto113	China	1320C	1763
98	114	Producto114	Francia	1692C	545
99	115	Producto115	India	1395I	1292

5 COLUMNAS, 99 FILAS Generación de perfiles de columnas basada en las 1000 primeras filas

Como orientación, los pasos seguidos en este proceso ETL son

1. Abrimos una nueva hoja para vincular los datos originales.
2. FASE TRANSFORMACION y LIMPIEZA
 - 2.1. Así el primer paso es crear una conexión a la fuente a través de Nueva Consulta. Iniciamos el editor Power Query (Datos – Obtener Datos) y seleccionamos nuevo origen de datos y que en este caso es un fichero CSV al cual creamos la conexión..

- 2.2. Confirmamos que se ha leído correctamente los encabezados de los campos y en caso contrario “Encabezados Promovidos”, con la opción Usar la Primera fila como encabezado. Así la tabla ha de tener los encabezados que tiene en la primera fila.
- 2.3. Para quitar las filas vacías, seleccionamos la columna Producto ID y activamos el filtro seleccionando la opción “Quitar Vacíos”.

Table with columns: Product ID, Producto, Made in, Batch code, Quantity. The 'Producto ID' column is filtered to show only rows with values from 'Producto1' to 'Producto10'.

Product ID	Producto	Made in	Batch code	Quantity
Producto1	China	1220C	370	
Producto2	Francia	1314C	2884	
Producto3	India	1996I	2248	
Producto4	Turquia	1348P	2878	
Producto5	Reino Unido	1904U	3201	
Producto6	India	1030I	757	
Producto7	Turquia	1007P	2435	
Producto8	Reino Unido	1040U	1955	
Producto9	Alemania	1987A	3049	
Producto10	China	1744C	1260	

- 2.4. Igualmente nos interesa quitar aquellos registros que contienen NA, para ello como hicimos en el paso anterior activamos el filtro de la columna Producto ID y deseleccionamos la categoría NA.

Table with columns: Product ID, Producto, Made in, Batch code, Quantity. The 'Producto ID' column is filtered to show rows with values from 'Producto15' to 'Producto19' and 'Producto30', with 'NA' entries deselected.

Product ID	Producto	Made in	Batch code	Quantity
Producto15	Francia	1644C	343	
Producto16	India	1428I	2175	
Producto17	Turquia	1803P	3800	
Producto18	Reino Unido	1143U	1153	
Producto19	India	1334I	3009	
NA	NA	NA	NA	
NA	NA	NA	NA	
NA	NA	NA	NA	
NA	NA	NA	NA	
NA	NA	NA	NA	
NA	NA	NA	NA	
NA	NA	NA	NA	
Producto30	India	1059I	2142	

- 2.5. Revisamos y cambiamos el tipo de dato que tiene cada columna para ello simplemente con el icono de la izquierda del nombre del campo podemos hacerlo. Así por ejemplo la columna Quantity esta como texto y nos interesa cambiar a numérico.
- 2.6. Así tenemos la tabla normalizada con 99 registros.

Resumen de los pasos seguidos en el proceso ETL

Configuración de la co... x

- PROPIEDADES**
 - Nombre: 103 Original Quantity
 - Todas las propiedades
- PASOS APLICADOS**
 - Origen *
 - Encabezados promovidos *
 - Filas filtradas *
 - X Tipo cambiado ***

Conceptos y técnicas requeridas

Los conceptos y técnicas para aplicar con Power Query relacionadas con el proceso ETL son:

Creación de una consulta, conexión a un fichero fuente no normalizado tipo CSV.

Uso de las opciones “Encabezados Promovidos” opción Usar la Primera fila como encabezado

Aplicamos filtros para no disponer de determinadas filas, por ejemplo, que contiene NA en una determinada columna. Quitar vacíos

Cambiar tipo de datos de las columnas a través del icono izquierdo, pasamos de texto a número entero.