Laboratorio #5 – ETL con manejo de historia

Andrés M. Ochoa Toro (201913554)
Manuel Felipe Porras Tascón(201913554)
Fernando Andrés Ávalos López(201911468)
ISIS 3710 – Inteligencia de Negocios
Universidad de los Andes, Bogotá, Colombia
{am.ochoat, fa.avalos, mf.porras}@uniandes.edu.co
Fecha de presentación: Noviembre 20 de 2021

Tabla de contenido

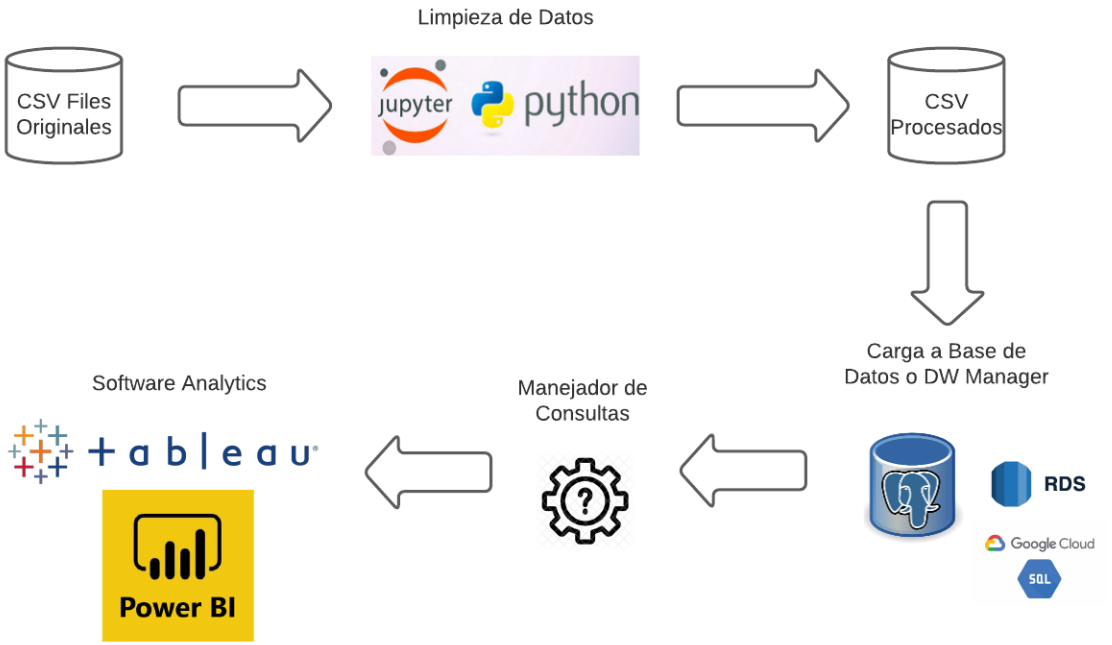
1 Introducción

WWI (World Wide Importers) es una empresa encargada de realizar importaciones y venderlas a diferentes clientes en diferentes ciudades de Estados Unidos. En esta ocasión, WWI desea optimizar sus ganancias, pues consideran que algunos de sus productos no están generando las ganancias que deberían. Para esta primera fase, se desea realizar la creación de la base de datos, la carga de datos y la comprobación del correcto funcionamiento del proceso realizado a través de algunas consultas iniciales.

En esta etapa de la consultoría, la empresa desea implementar el proceso ETL que le permita extraer los datos de órdenes desde unos archivos CSV y almacenarlos en un modelo dimensional tal que les permita realizar análisis para mejorar entre otros elementos, su eficiencia operativa. A continuación, se presenta el modelo multidimensional que se desea obtener:

2 Perfilamiento y procesamiento de datos

El proceso ETL realizado en este laboratorio, se describe en el siguiente diagrama ETL.



3 Ventajas y desventajas de estrategias del manejo

Las ventajas y desventajas de las maneras propuestas para el manejo de historias se describen a continuación.

Estrategia	Desventajas	Ventajas
Tipo 1	Esta estrategia olvida toda acción que se haya realizado, lo cual no permite realizar consultas o análisis con datos anteriores a los que se tienen actualmente. Esto presenta una dificultad para los análisis que necesite la empresa sobre sus acciones.	Los recursos respecto a la capacidad de almacenamiento no aumentan a medida de las actualizaciones realizadas. Esto permite que los recursos de almacenamiento necesarios por la empresa no representen un gran gasto. Así mismo, representa un esquema bastante simple y fácil de operar.
Tipo 2	Esta estrategia representa una gran	En un conjunto de datos, que tenga un

	complejidad respecto a la capacidad de almacenamiento. Si se tiene un conjunto significativo de datos que sean actualizados muy frecuentemente, el número de filas crecerá muy rápidamente por la necesidad de conservar los cambios.	poco número de filas y sean actualizadas con poca frecuencia, se permite la posibilidad de conservar un gran historial de datos. Esto genera que al momento de realizar un análisis profundo del negocio y los datos que lo componen.
Tipo 3	Para datos que cambian impredeciblemente representa un esquema poco eficiente, ya que sería necesario la creación de una gran cantidad de columnas con valores nulos.	Para un conjunto de datos en el cual sea predecible las columnas que van a cambiar, representa una forma eficiente de a nivel de almacenamiento, esto porque permite establecer una columna que capture los cambios. Así mismo representa una forma para análisis puntuales en los cambios que se pueden necesitar.

4 Perfilamiento y procesamiento de datos

Para este caso, se usaron los mismos datos que se obtuvieron como resultado del laboratorio pasado. En el caso de los nuevos datos que se proporcionan, se hace un nuevo proceso de limpieza de datos para eliminar los valores nulos, los cuales pueden representar un problema en su manejo en la base de datos.

Nota Importante: Todas la información, archivos, datos y notebooks se encuentran en el siguiente repositorio: https://github.com/AmOchoat/BI_Laboratorio5 .

5 Herramientas proceso ETL

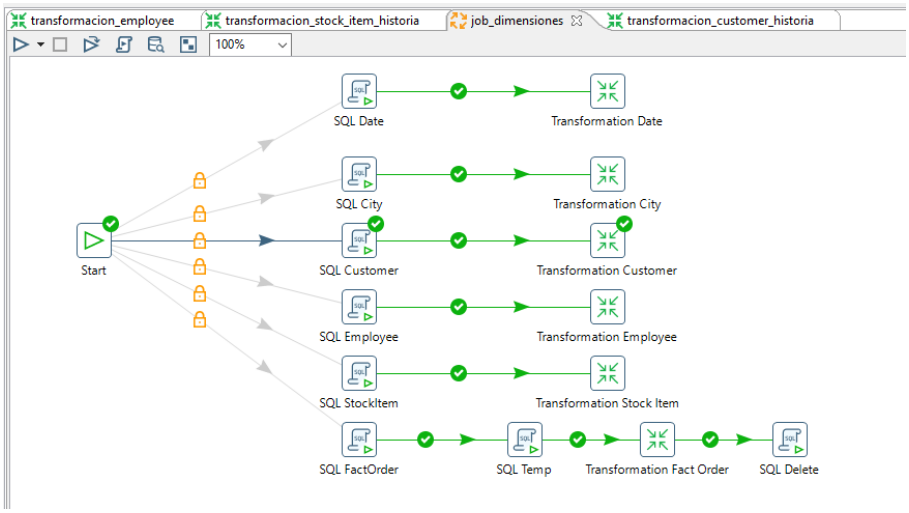
Para el proceso de ETL, equivalente al laboratorio 5, se continúa usando la base de datos que se creó en el laboratorio 4, y se añade el uso de las siguientes herramientas.

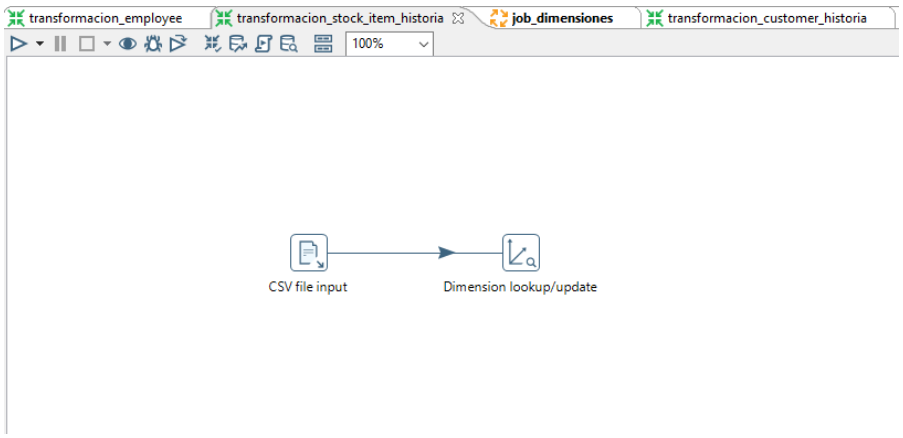
Nota Importante: Todas la información, archivos, datos y notebooks se encuentran en el siguiente repositorio: https://github.com/AmOchoat/BI_Laboratorio5 .

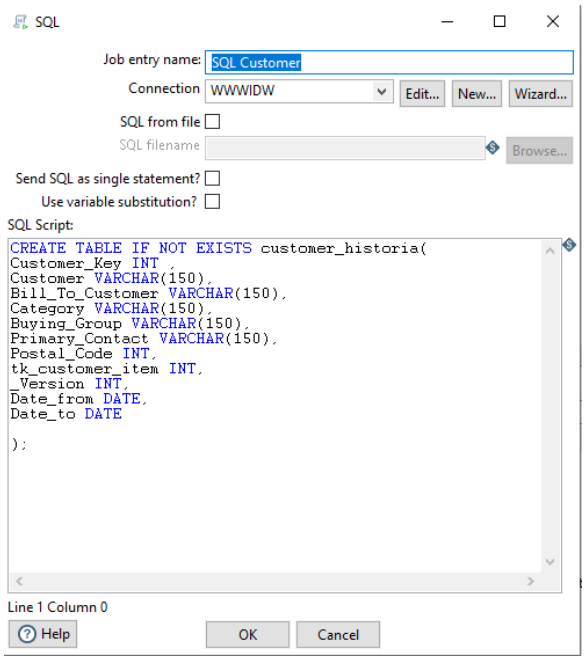
5.1 Spoon

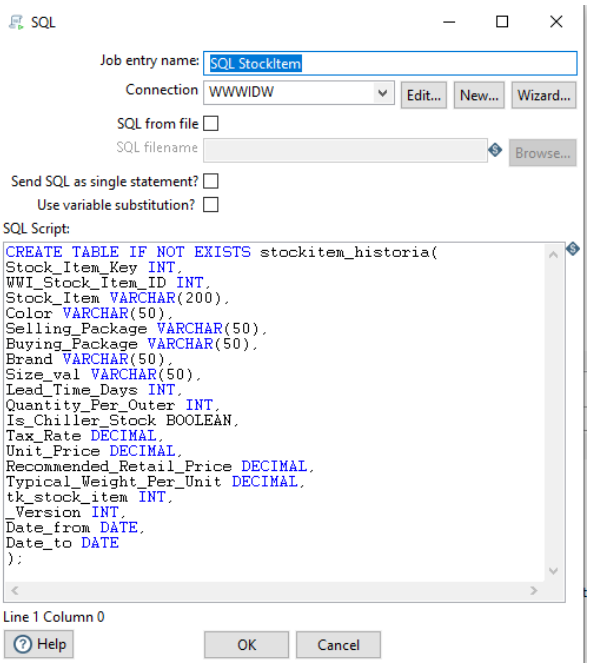
Haciendo uso de la herramienta Spoon, se realizó el siguiente procedimiento ETL.

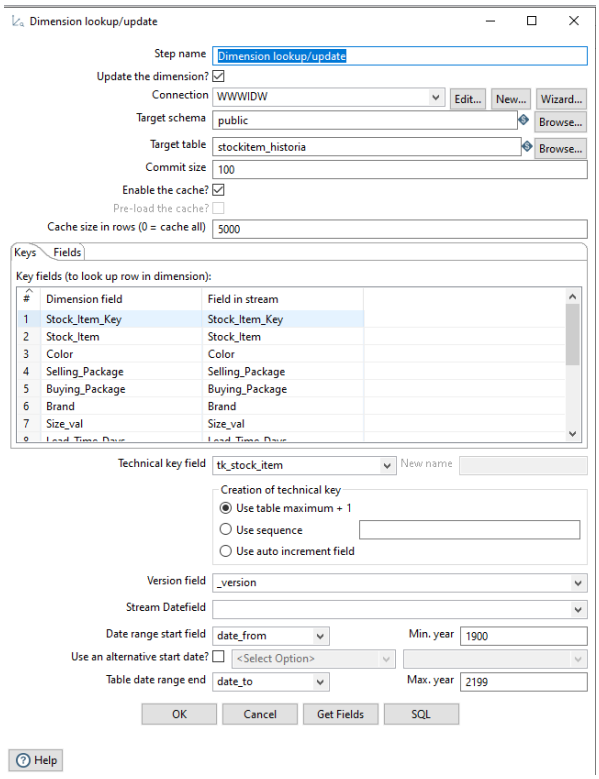
En primer lugar, se implemento el tipo 2 de Kimball tanto para la dimensión pedida en la guía del laboratorio que fue *stock_item* como para la dimensión *customer* que es pedida como entregable. Además, se modificó la tabla de hechos como era pedida y como era necesario. Se adjuntan las pruebas correspodientes de las implementaciones:

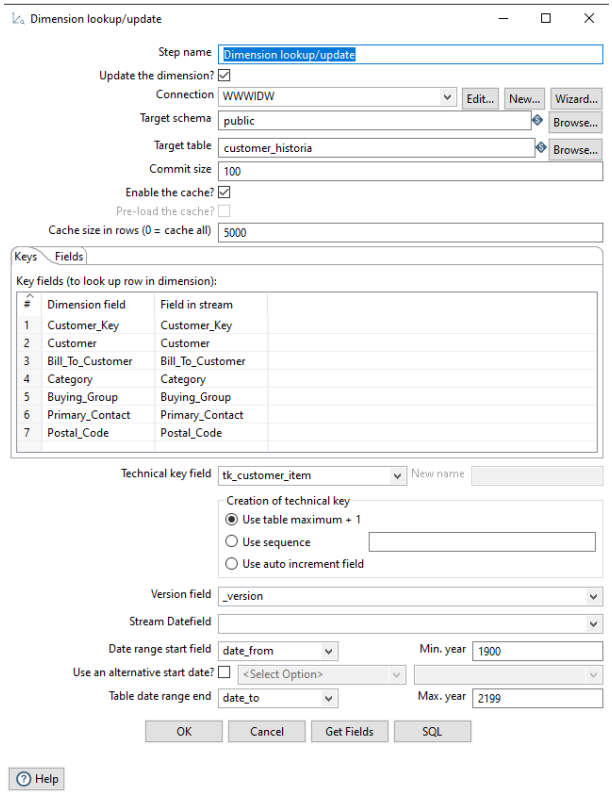





SQL
Job entry name:
SQL Customer
Connection
WWWIDW
Edit...
New...
Wizard...
SQL from file
SQL filename
Browse...
Send SQL as single statement?
Use variable substitution?
SQL Script:
CREATE TABLE IF NOT EXISTS customer_historia(
Customer_Key INT,
Customer VARCHAR(150),
Bill_To_Customer VARCHAR(150),
Category VARCHAR(150),
Buying_Group VARCHAR(150),
Primary_Contact VARCHAR(150),
Postal_Code INT,
tk_customer_item INT,
_Version INT,
Date_from DATE,
Date_to DATE
);
Line 1 Column 0
Help
OK
Cancel


SQL
Job entry name:
SQL StockItem
Connection
WWWIDW
Edit...
New...
Wizard...
SQL from file
SQL filename
Browse...
Send SQL as single statement?
Use variable substitution?
SQL Script:
CREATE TABLE IF NOT EXISTS stockitem_historia(
Stock_Item_Key INT,
WWI_Stock_Item_ID INT,
Stock_Item VARCHAR(200),
Color VARCHAR(50),
Selling_Package VARCHAR(50),
Buying_Package VARCHAR(50),
Brand VARCHAR(50),
Size_val VARCHAR(50),
Lead_Time_Days INT,
Quantity_Per_Outer INT,
Is_Chiller_Stock BOOLEAN,
Tax_Rate DECIMAL,
Unit_Price DECIMAL,
Recommended_Retail_Price DECIMAL,
Typical_Weight_Per_Unit DECIMAL,
tk_stock_item INT,
_Version INT,
Date_from DATE,
Date_to DATE
);
Line 1 Column 0
Help
OK
Cancel





Se comprueba a través de pgAdmin la creación de las tablas y la carga de datos:

WWWIDW/postgres@PostgreSQL 14

Query Editor | Query History

```
SELECT order_key, city_key, customer_key, stock_item_key, order_date_key, picked_date_key, salesperson_key, picker_key, "package", quant
    FROM public.fact_order;
```

Data Output | Explain | Messages | Notifications

	order_key [PK] integer	city_key integer	customer_key integer	stock_item_key integer	order_date_key date	picked_date_key date	salesperson_key integer	picker_key integer	package character varying (50)	quantity integer	unit_p numer
1	2	83	2	631	2015-11-17	2013-07-19	2	133	S	76	
2	23	68	221	215	2014-05-16	2014-05-06	54	156	S	919	
3	24	51	382	399	2016-06-18	2016-02-26	186	51	S	589	
4	25	82	78	336	2014-09-08	2013-01-21	74	188	L	92	
5	26	19	19	15	2016-01-16	2014-04-20	87	191	S	180	
6	27	70	182	366	2016-08-05	2014-01-05	105	201	XL	873	
7	28	39	342	499	2013-12-13	2016-10-21	185	153	S	275	
8	29	75	115	292	2013-06-21	2014-05-16	76	179	S	627	
9	30	13	347	61	2014-12-22	2016-10-31	147	88	M	726	
10	31	50	122	13	2015-08-26	2016-02-27	136	156	S	245	
11	32	46	144	372	2014-06-16	2013-05-12	187	146	S	534	
12	33	34	65	351	2014-12-30	2016-10-10	94	162	S	944	
13	34	79	13	637	2016-06-14	2014-02-27	33	168	XL	120	
14	35	67	348	192	2016-02-18	2015-05-01	122	57	M	802	
15	36	82	235	267	2014-06-26	2015-03-14	103	86	S	337	
16	37	61	289	428	2014-10-15	2013-01-20	108	189	M	590	
17	38	20	255	332	2015-11-03	2014-03-21					

Successfully run. Total query runtime: 106 msec. 1030 rows affected.

WWWIDW/postgres@PostgreSQL 14

Query Editor | Query History | Scratch Pad

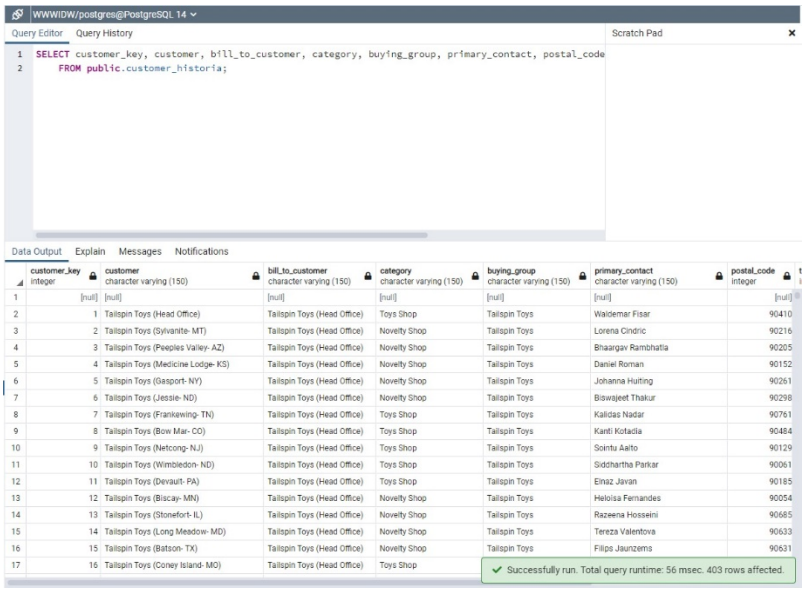
```
SELECT is_chiller_stock, tk_stock_item, _version, date_from, date_to, stock_item_key, stock_item, col
    FROM public.stockitem_historia;
```

Data Output | Explain | Messages | Notifications

	is_chiller_stock boolean	tk_stock_item integer	_version integer	date_from date	date_to date	stock_item_key integer	stock_item character varying (200)	color character v
1	false	426	1	1900-01-01	2199-12-31	425	Alien officer hoodie (Black) XL	Black
2	[null]	0	1	[null]	[null]	[null]	[null]	[null]
3	false	1	1	1900-01-01	2199-12-31	0	Unknown	N/A
4	false	2	1	1900-01-01	2199-12-31	1	Void fill 400 L bag (White) 400L	N/A
5	false	3	1	1900-01-01	2199-12-31	2	Void fill 300 L bag (White) 300L	N/A
6	false	4	1	1900-01-01	2199-12-31	3	Void fill 200 L bag (White) 200L	N/A
7	false	5	1	1900-01-01	2199-12-31	4	Void fill 100 L bag (White) 100L	N/A
8	false	6	1	1900-01-01	2199-12-31	5	Air cushion machine (Blue)	N/A
9	false	21	1	1900-01-01	2199-12-31	20	Black and yellow heavy despatch tape 48mmx100m	N/A
10	false	22	1	1900-01-01	2199-12-31	21	Black and yellow heavy despatch tape 48mmx75m	N/A
11	false	23	1	1900-01-01	2199-12-31	22	Black and orange this way up despatch tape 48mmx100m	N/A
12	false	24	1	1900-01-01	2199-12-31	23	Black and orange this way up despatch tape 48mmx75m	N/A
13	false	25	1	1900-01-01	2199-12-31	24	Black and orange handle with care despatch tape 48mmx100m	N/A
14	false	242	1	1900-01-01	2199-12-31	241	USB food flash drive - cookie	N/A
15	false	337	1	1900-01-01	2199-12-31	336	Bubblewrap dispenser (Red) 1.5m	Red
16	false	36	1	1900-01-01	2199-12-31	35	Shipping carton (Brown) 356x356x279mm	N/A
17	false	37	1	1900-01-01	2199-12-31	36	Shipping cart	

Read-only column

Successfully run. Total query runtime: 65 msec. 781 rows affected.



Preguntas:

- ¿Cuál es el objetivo de la columna *tk_stock_item*?

  El objetivo de esta columna es tener una referencia con los registros de los objetos a vender ya que el objetivo de tener otra tabla llamada *StockItem_historia* es llevar un historial y registro de los productos como lo sería una implementación de un diagrama multidimensional de tipo 2 propuesto por Kimball.

  - ¿Qué significa cada una de estas opciones?

  Básicamente, la descripción de cada una de las opciones es la siguiente:

  - Technical key: Es la lalve subrogada que tiene auto incremento y es única.

  - Version Field: Guarda la versión de la entrada de la dimensión. Ejemplo: 1 ,2, 3 …

  - Start of date range: Indica la fecha de inicio de la versión.

  - End of date range: Indica la fecha de finalización de la versión.

  - Lookup/ Update Fields: Acá existen varias opciones con respecto a los campos que se van a actualizar o se va accionar uno de los siguientes comandos sobre ese campo:

  - Insert: Esta opción implementa un tipo II lentamente cambiando la dimensión política. Si la diferencia es detectada

para una o más mapeos que tienen la opción de Insert, luego una fila es añadida a la tabla de dimensión.

- Update: Esta opción simplemente actualiza las filas encontradas. Además, este puede usarse para implementar un Tipo I cambiando lentamente la dimensión.

- Punch through: La opcion The punch through implementa una actualización , pero en vez de solo actualizar las respectivas filas , este actualizará todas las versiones de la fila en un tipo 2.

- Date of last insert or update (without stream field as source): Esta opción deja el paso mantener un registro de fecha que guarda la fecha de inserción o actualización automáticamente con la fecha del sistema.

- Date of last insert (without stream field as source): Esta opción deja el paso mantener un registro de fecha que guarda la fecha de inserción automáticamente con la fecha del sistema.

- Date of last update (without stream field as source): Esta opción deja el paso mantener un registro de fecha que guarda la fecha de actualización automáticamente con la fecha del sistema.

- Last version (without stream field as source): Esta opción deja el paso de mantener un flag para indicar si la fila es la última version , todo automáticamente.

- ¿Cómo se puede saber que el proceso ETL manejó los cambios entre los dos archivos?` Pista: Puede revisar los valores de las columnas creadas para el manejo de historia.

  Se puede entender que se manejaron los cambios entre los dos archivos, ya que se debieron haber creado otros registros con ligeros cambios en la información y referenciados como _version 2, así indicando que hubo un cambio y se generó otro registro para obtener el registro actual manteniendo el registro pasado.

- ¿Por qué se debe hacer el cambio para que la columna de la llave subrogada sea de tipo *Unique*?

  Se hace para que sea único esa llave subrogada, ya que esa llave no tiene ningún valor para el negocio, pero nos permite tener un identificador en los registrados y así estos estén diferenciados.

- ¿Qué pasa si uno de los datos reportados en la tabla de hechos no existe en alguna de las dimensiones?

  No tendría sentido, ya que la tabla de hechos en su mayoría tiene llaves foráneas por lo que generaría un error.
- ¿Qué sugiere para evitar esa situación?

  Para evitar esta situación, se sugiere que se limpien los datos dado que se hago un mapeo entre los datos de la tabla de hechos y ver si cada una de las llaves foráneas existen y de lo contrario no tomar en cuenta esos datos.

5.2 PostgreSQL y Python

Como segundo conjunto de herramientas, se combina PostgreSQL para representar el *Data Warehouse* y Python para realizar el debido proceso.

Para este caso, primero se procede a crear una conexión con la base de datos existente. Posteriormente se procede a cargar los datos actuales de la base de datos. Este primer proceso de carga se ve acompañado con la creación de nuevas columnas para el manejo de historia tipo 2. Estas columnas van a representar la versión, la fecha de inicio y fin de la versión, y la nueva llave para el *stock_item*. Estos nuevos datos se llenan con valores por defecto, sea la versión uno o la máxima y mínima fecha soportada

Para el manejo de los nuevos datos, se procede a cargar el archivo en un nuevo *dataframe,* al cual se le realiza un proceso de limpieza para eliminar los valores nulos que existan o en caso extremo eliminar toda una columna, como se hizo con la columna *Brand.* Después de este paso, se procede a procesar los dos conjuntos de datos de tal manera que se comprueba la existencia de la llave *stock_item_key* en ambos conjuntos de datos. Si esto se da, se procede a comparar cada una de las columnas para ver si hay alguna actualización, y proceder a realizar a realizar el respectivo manejo de historias.

El respectivo resultado de este proceso se vería en la base de datos con las versiones que representan la actualización de los productos en un rango de fecha. Lo cual se observa en la siguiente imagen.

	stock_item_key bigint	stock_item text	color text	selling_package text	buying_package text	size_val text	lead_time_days bigint	quantity_per_outer bigint	is_chiller_stock boolean	tax_rate text	unit_price text	recommended_retail_price double precision	typical_w double pr
1	1	Void fill 400 L bag (White) 400L	Blue	Each	Each	400L	14	10	false	14.000	50.00	74.75	
2	1	Void fill 400 L bag (White) 400L	Unknown	Each	Each	400L	14	10	false	14.0	50.0	74.75	
3	2	Void fill 300 L bag (White) 300L	Blue	Each	Each	300L	14	10	false	14.000	37.50	56.06	
4	2	Void fill 300 L bag (White) 300L	Unknown	Each	Each	300L	14	10	false	14.0	37.5	56.06	
5	3	Void fill 200 L bag (White) 200L	Unknown	Each	Each	200L	14	10	false	14.0	25.0	37.38	
6	3	Void fill 200 L bag (White) 200L	Blue	Each	Each	200L	14	10	false	14.000	25.00	37.38	
7	4	Void fill 100 L bag (White) 100L	Unknown	Each	Each	100L	14	10	false	14.0	12.5	18.69	
8	4	Void fill 100 L bag (White) 100L	Unknown	Each	Each	100L	14	10	false	14.000	12.50	18.69	
9	5	Air cushion machine (Blue)	Unknown	Each	Each	Unknown	20	1	false	20.0	1899.0	2839.01	
10	5	Air cushion machine (Blue)	Gray	Each	Each	Unknown	20	1	false	20.000	1899.00	2839.01	
11	6	Air cushion film 200mmx200mm 325m	Unknown	Each	Each	325m	14	1	false	14.0	90.0	134.55	
12	6	Air cushion film 200mmx200mm 325m	Unknown	Each	Each	325m	14	1	false	14.000	90.00	134.55	
13	7	Air cushion film 200mmx100mm 325m	Unknown	Each	Each	325m	14	1	false	14.0	87.0	130.07	
14	7	Air cushion film 200mmx100mm 325m	Unknown	Each	Each	325m	14	1	false	14.000	87.00	130.07	
15	8	Large replacement blades 18mm	Unknown	Each	Each	18mm	14	10	false	14.0	4.3	6.43	
16	8	Large replacement blades 18mm	Unknown	Each	Each	18mm	14	10	false	14.000	4.30	6.43	
17	9	Small 9mm replacement blades 9mm	Unknown	Each	Each	9mm	14	10	false	14.0	4.1	6.13	
18	9	Small 9mm replacement blades 9mm	Unknown	Each	Each	9mm	14	10	false	14.000	4.10	6.13	
19	10	Packing knife with metal insert blade (Yellow) 18mm	Unknown	Each	Each	18mm	14	5	false	14.000	2.40	3.59	
20	10	Packing knife with metal insert blade (Yellow) 18mm	Unknown	Each	Each	18mm	14	5	false	14.0	2.4	3.59	
21	11	Packing knife with metal insert blade (Yellow) 9mm	Unknown	Each	Each	9mm	14	5	false	14.0	1.89	2.83	
22	11	Packing knife with metal insert blade (Yellow) 9mm	Blue	Each	Each	9mm	14	5	false	14.000	1.89	2.83	
23	12	Permanent marker red 5mm nib (Red) 5mm	Unknown	Each	Each	5mm	14	12	false	14.0	2.7	4.04	
24	12	Permanent marker red 5mm nib (Red) 5mm	Blue	Each	Each	5mm	14	12	false	14.000	2.70	4.04	
25	13	Permanent marker blue 5mm nib (Blue) 5mm	Blue	Each	Each	5mm	14	12	false	14.000	2.70	4.04	
26	13	Permanent marker blue 5mm nib (Blue) 5mm	Unknown	Each	Each	5mm	14	12	false	14.0	2.7	4.04	
27	14	Permanent marker black 5mm nib (Black) 5mm	Blue	Each	Each	5mm	14	12	false	14.000	2.70	4.04	
28	14	Permanent marker black 5mm nib (Black) 5mm	Unknown	Each	Each	5mm	14	12	false	14.0	2.7	4.04	

Resultados manejo de historias stock item 1

Data Output | Explain | Messages | Notifications

size_val text	lead_time_days bigint	quantity_per_outer bigint	is_chiller_stock boolean	tax_rate text	unit_price text	recommended_retail_price double precision	typical_weight_per_unit double precision	tk_stock_item bigint	_version bigint	date_from timestamp without time zone	date_to timestamp without time zone	recommended_retail_Price text	typical_Weight_Per_Unit text
400L	14	10	false	14.000	50.00	74.75	1	11408	2	2021-11-20 00:00:00	2199-12-31 00:00:00	74.75	1.000
400L	14	10	false	14.0	50.0	74.75	1	0	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
300L	14	10	false	14.000	37.50	56.06	0.75	12769	2	2021-11-20 00:00:00	2199-12-31 00:00:00	56.06	.750
300L	14	10	false	14.0	37.5	56.06	0.75	1	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
200L	14	10	false	14.0	25.0	37.38	0.5	2	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
200L	14	10	false	14.000	25.00	37.38	0.5	12788	2	2021-11-20 00:00:00	2199-12-31 00:00:00	37.38	.500
100L	14	10	false	14.0	12.5	18.69	0.25	3	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
100L	14	10	false	14.000	12.50	18.69	0.25	12807	2	2021-11-20 00:00:00	2199-12-31 00:00:00	18.69	.250
Unknown	20	1	false	20.0	1899.0	2839.01	10	4	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
Unknown	20	1	false	20.000	1899.00	2839.01	10	12826	2	2021-11-20 00:00:00	2199-12-31 00:00:00	2839.01	10.000
325m	14	1	false	14.0	90.0	134.55	6	5	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
325m	14	1	false	14.000	90.00	134.55	6	12845	2	2021-11-20 00:00:00	2199-12-31 00:00:00	134.55	6.000
325m	14	1	false	14.0	87.0	130.07	5	6	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
325m	14	1	false	14.000	87.00	130.07	5	12864	2	2021-11-20 00:00:00	2199-12-31 00:00:00	130.07	5.000
18mm	14	10	false	14.0	4.3	6.43	0.8	7	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
18mm	14	10	false	14.000	4.30	6.43	0.8	12883	2	2021-11-20 00:00:00	2199-12-31 00:00:00	6.43	.800
9mm	14	10	false	14.0	4.1	6.13	0.7	8	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
9mm	14	10	false	14.000	4.10	6.13	0.7	12902	2	2021-11-20 00:00:00	2199-12-31 00:00:00	6.13	.700
18mm	14	5	false	14.000	2.40	3.59	0.5	12921	2	2021-11-20 00:00:00	2199-12-31 00:00:00	3.59	.500
18mm	14	5	false	14.0	2.4	3.59	0.5	9	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
9mm	14	5	false	14.0	1.89	2.83	0.5	10	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
9mm	14	5	false	14.000	1.89	2.83	0.5	12940	2	2021-11-20 00:00:00	2199-12-31 00:00:00	2.83	.500
5mm	14	12	false	14.0	2.7	4.04	0.1	11	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
5mm	14	12	false	14.000	2.70	4.04	0.1	12959	2	2021-11-20 00:00:00	2199-12-31 00:00:00	4.04	.100
5mm	14	12	false	14.000	2.70	4.04	0.1	12978	2	2021-11-20 00:00:00	2199-12-31 00:00:00	4.04	.100
5mm	14	12	false	14.0	2.7	4.04	0.1	12	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]
5mm	14	12	false	14.000	2.70	4.04	0.1	12997	2	2021-11-20 00:00:00	2199-12-31 00:00:00	4.04	.100
5mm	14	12	false	14.0	2.7	4.04	0.1	13	1	1900-01-01 00:00:00	2021-11-20 00:00:00	[null]	[null]

Resultados manejo de historias stock item 2

Como siguiente paso, se procede a procesar la tabla de hechos. En este caso se realiza un proceso similar a lo mencionado para *stock_item*. Se parte de la carga de los conjuntos de datos a un *dataframe* para posteriormente realizar la creación del nuevo datafreme que va a ser guardado en la base de datos.

El proceso realizado en este caso es comparar la llave *stock_item_key,* existente en el conjunto de datos *fact_order*, para buscarla en *dataframe* correspondiente a *stock_item*. Si esta llave se encuentra se procede a comparar la fecha en que se realizó la orden en el rango que abarca la versión del producto en *stock_item*. El resultado de este es la actualización de *stock_item_key* con el valor de *tk_stock_item* de la fila encontrada.

Data Output | Explain | Messages | Notifications

	city_key bigint	customer_key bigint	stock_item_key bigint	order_date_key text	picked_date_key text	salesperson_key bigint	picker_key bigint	package text	quantity bigint	unit_price double precision	tax_rate bigint	total_excluding_tax double precision	tax_amount double precision	total_including_tax double precision
1	43	8	0	2014-07-01	2014-11-17	30	14	Bag	38	6107.41	69	21644.66	36567.51	202482.64
2	66	10	0	2014-07-01	2014-11-17	119	125	Big Box	70	6910.22	53	17980.76	31883.82	108897.62
3	2	8	0	2014-07-01	2014-11-17	70	132	Packed	54	2391.04	46	38043.91	4200.09	262277.55
4	92	10	0	2014-07-01	2014-11-17	175	194	Packed	27	3674.87	51	78158.19	26664.43	462262.27
5	92	202	1	2014-09-19	2014-08-26	51	39	X	783	4534.8	67	8727.83	103.62	2696.54
6	57	320	1	2016-11-07	2013-08-16	39	16	S	515	1081.3	58	3082.19	238.35	8127.04
7	41	194	3	2013-06-19	2013-07-19	22	80	XL	166	3099.48	67	4251.39	526.07	2403.81
8	59	107	3	2015-09-25	2016-02-13	49	152	S	571	2275.16	40	7236.11	66.38	2640.81
9	35	56	3	2016-11-21	2015-09-25	152	2	S	913	1517.48	31	7980.79	602.9	6003.26
10	59	275	3	2015-10-07	2014-01-10	201	126	S	368	2449.78	3	8102.71	333.49	2564.07
11	45	69	3	2016-01-02	2013-10-25	33	151	XL	247	1125.7	58	1863.61	430.32	7416.21
12	52	370	4	2015-03-20	2014-11-02	11	113	L	557	4148	67	6848.27	416.5	6379.57
13	19	365	5	2013-01-09	2013-09-04	9	192	S	312	3800.47	9	5131.55	910.51	4207.29
14	10	299	5	2015-10-22	2014-09-08	155	48	S	316	1681.68	23	915.21	537.36	1174.39
15	7	101	6	2015-12-19	2014-02-26	201	108	S	5	3124.77	19	1762.18	613.16	7663.45
16	14	183	6	2014-07-23	2014-06-01	120	140	M	383	4539.9	10	2616.44	674.83	1002.25
17	74	103	6	2016-08-31	2013-06-29	69	155	S	783	3609.21	64	6263.25	629.72	7416.58
18	85	50	8	2014-06-12	2016-09-29	29	143	M	512	3577.56	53	1988.87	642.65	8100.82
19	57	217	8	2014-09-07	2014-05-24	51	209	S	1	1524.38	65	770.58	360.63	1561.39
20	36	140	9	2016-05-28	2014-03-27	88	149	S	286	804.29	4	484.45	553.62	7784.35
21	94	323	10	2014-06-30	2015-07-24	109	209	L	323	3062.77	1	8782.39	53.93	4323.88

Resultado fact_order

Para el caso del conjunto de datos *customer*. Se aplicó el proceso para los tres tipos de historias, siendo esto el tipo 1, 2, 3. A continuación se refleja los resultados obtenidos a través de consultas SQL en PostgreSQL.

Data Output | Explain | Messages | Notifications

	Customer text	Bill_To_Customer text	Category text	Buying_Group text	Primary_Contact text	Postal_Code double precision
1	Tailspin Toys (Head Office)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Waldemar Fisar	90410
2	Tailspin Toys (Sylvanite- MT)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Lorena Cindric	90216
3	Tailspin Toys (Peeples Valley- AZ)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Bhaargav Rambhatla	90205
4	Tailspin Toys (Medicine Lodge- KS)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Daniel Roman	90152
5	Tailspin Toys (Gasport- NY)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Johanna Huiting	90261
6	Tailspin Toys (Jessie- ND)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Biswajeet Thakur	90298
7	Tailspin Toys (Frankewing- TN)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Kalidas Nadar	90761
8	Tailspin Toys (Bow Mar- CO)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Kanti Kotadia	90484
9	Tailspin Toys (Netcong- NJ)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Sointu Aalto	90129
10	Tailspin Toys (Wimbledon- ND)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Siddhartha Parkar	90061
11	Tailspin Toys (Devault- PA)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Elnaz Javan	90185
12	Tailspin Toys (Biscay- MN)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Heloisa Fernandes	90054
13	Tailspin Toys (Stonefort- IL)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Razeena Hosseini	90685
14	Tailspin Toys (Long Meadow- MD)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Tereza Valentova	90633
15	Tailspin Toys (Batson- TX)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Filips Jaunzems	90631
16	Tailspin Toys (Coney Island- MO)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Nitin Matondkar	90467
17	Tailspin Toys (East Fultonham- OH)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Adam Kubat	90416
18	Tailspin Toys (Goffstown- NH)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Isabelle Vodlan	90321
19	Tailspin Toys (Lemeta- AK)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Mithun Bhattacharya	90303
20	Tailspin Toys (College Place- WA)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Nghi Hua	90050
21	Tailspin Toys (Tresckow- PA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Duleep Raju	90790
22	Tailspin Toys (Ward Ridge- FL)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Cristina Longo	90784
23	Tailspin Toys (Ikatan- AK)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Hang Tang	90019
24	Tailspin Toys (Dundarrach- NC)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Intira Mookjai	90758
25	Tailspin Toys (Avenal- CA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Sulabha Khalsa	90352
26	Tailspin Toys (Hedrick- IA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Dhanishta Majji	90331

Resultado Customer Tipo 1

Data Output | Explain | Messages | Notifications

	Customer text	Bill_To_Customer text	Category text	Buying_Group text	Primary_Contact text	Postal_Code double precision	_version bigint
1	Tailspin Toys (Head Office)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Waldemar Fisar	90410	1
2	Tailspin Toys (Sylvanite- MT)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Lorena Cindric	90216	1
3	Tailspin Toys (Peeples Valley- AZ)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Bhaargav Rambhatla	90205	1
4	Tailspin Toys (Medicine Lodge- KS)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Daniel Roman	90152	1
5	Tailspin Toys (Gasport- NY)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Johanna Huiting	90261	1
6	Tailspin Toys (Jessie- ND)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Biswajeet Thakur	90298	1
7	Tailspin Toys (Frankewing- TN)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Kalidas Nadar	90761	1
8	Tailspin Toys (Bow Mar- CO)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Kanti Kotadia	90484	1
9	Tailspin Toys (Netcong- NJ)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Sointu Aalto	90129	1
10	Tailspin Toys (Wimbledon- ND)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Siddhartha Parkar	90061	1
11	Tailspin Toys (Devault- PA)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Elnaz Javan	90185	1
12	Tailspin Toys (Biscay- MN)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Heloisa Fernandes	90054	1
13	Tailspin Toys (Stonefort- IL)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Razeena Hosseini	90685	1
14	Tailspin Toys (Long Meadow- MD)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Tereza Valentova	90633	1
15	Tailspin Toys (Batson- TX)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Filips Jaunzems	90631	1
16	Tailspin Toys (Coney Island- MO)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Nitin Matondkar	90467	1
17	Tailspin Toys (East Fultonham- OH)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Adam Kubat	90416	1
18	Tailspin Toys (Goffstown- NH)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Isabelle Vodlan	90321	1
19	Tailspin Toys (Lemeta- AK)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Mithun Bhattacharya	90303	1
20	Tailspin Toys (College Place- WA)	Tailspin Toys (Head Office)	Toys Shop	Tailspin Toys	Nghi Hua	90050	1
21	Tailspin Toys (Tresckow- PA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Duleep Raju	90790	1
22	Tailspin Toys (Ward Ridge- FL)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Cristina Longo	90784	1
23	Tailspin Toys (Ikatan- AK)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Hang Tang	90019	1
24	Tailspin Toys (Dundarrach- NC)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Intira Mookjai	90758	1
25	Tailspin Toys (Avenal- CA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Sulabha Khalsa	90352	1
26	Tailspin Toys (Hedrick- IA)	Tailspin Toys (Head Office)	Novelty Shop	Tailspin Toys	Dhanishta Majji	90331	1

Resultado Customer Tipo 2

6 Recomendaciones

6.1 Hacer las transformaciones con una herramienta similar a SPOON es más natural y ameno que hacerlo con una herramienta como Pandas. No obstante Pandas es más configurable y versátil.

6.2 Sin importar la herramienta, es siempre perentorio realizar una limpieza de datos previa.

6.3 NO usar BigQuery puesto que no se podía modificar el esquema una vez cargados los datos.

6.4 Tener una conveción y exigir su cumplimiento en el nombramiento de las columnas de los datasets es recomendable puesto que aligera la tarea de SCD.