Preprocesamiento de datos sobre gestión de mantenimiento en una industria azucarera

Mauricio Ribas-García*
Instituto Cubano de Investigaciones de los Derivados de la Caña de Azúcar (ICIDCA)
Vía Blanca No. 804 y Carretera Central, San Miguel del Padrón. La Habana, Cuba.
* mauricio.ribas@icidca.azcuba.cu

RESUMEN

En los últimos años, el crecimiento masivo en la escala de los datos es un factor clave en el actual escenario de procesamiento de datos. La eficacia de los algoritmos para la extracción de conocimiento depende, en gran medida, de la calidad de los datos, esto puede ser garantizado con el empleo de su preprocesamiento. En este trabajo se exponen los resultados de la aplicación de un preprocesamiento a datos originales, almacenados en un libro Excel, que contiene información diaria de las salidas de materiales del almacén para tareas de mantenimiento, en un central azucarero. Se identificaron los errores en la estructura de almacenamiento de datos, desarrollada por el cliente, que dificultan la extracción del conocimiento. Se evidenciaron las ventajas que ofrece la herramienta *Power Query* de *Excel*, para realizar el preprocesamiento de datos y se construyó un reporte interactivo y automatizado que permite visualizar la cantidad de materiales de cada tipo extraído por cada área, durante el período analizado.

Palabras clave: calidad de datos, preprocesamiento de datos, transformación de datos, Power Query.

ABSTRACT

In recent years, the massive growth in data scale is becoming a key factor in the current data processing scenario. The effectiveness of the algorithms for the extraction of knowledge depend on data quality, this can be guaranteed with the use of data pre-processing. In this work, the results of the processing application to original data, stored in an Excel book, which contains daily information on the material output from the store for maintenance tasks in a sugar mill, are exposed. Errors in the data storage structure developed by the user that make it difficult to extract knowledge were identified. The advantages offered by the *Power Query* tool to perform data preprocessing were shown and an interactive and automated report was built that allows visualizing the amount of materials of each type extracted by maintenance area during the analyzed period. **Key words:** data quality, data preprocessing, data transformation, *Power Query*.

INTRODUCCIÓN

El preprocesamiento de datos es una etapa fundamental en el proceso de extracción de conocimiento, cuyo objetivo principal consiste en obtener un conjunto de datos de calidad para la extracción del conocimiento (1).

La calidad del conocimiento extraído depende, en gran medida, de la calidad de los datos. Los datos, en muchas ocasiones, están afectados por factores negativos como: ruido, valores perdidos, inconsistencias, valores superfluos o un tamaño demasiado grande en cualquier dimensión (número de atributos e instancias). Está demostrado que una baja calidad de los datos conduce, en la mayoría de los casos, a una baja calidad del conocimiento extraído (1).

En este artículo se exponen los resultados del preprocesamiento realizado a los datos originales, almacenados en un libro Excel, que contiene información diaria, durante el mes de abril de 2022, de las salidas de materiales, para las tareas de mantenimiento en un central azucarero.

El objetivo de este trabajo es presentar la importancia del preprocesamiento de datos y el uso de la herramienta *Power Query* de Excel para obtener, de forma automática, una base de datos de calidad que permita extraer conocimiento de los datos.

Para ello, el trabajo se organiza como sigue. En la sección 2 se realiza una breve descripción de la herramienta *Power Query* de Excel. En la sección 3 se exponen las principales inconsistencias encontradas en el fichero original de datos Excel. En la sección 4 se describen los diferentes pasos del preprocesamiento de datos aplicado. En la sección 5 se realiza la extracción del conocimiento a partir de los datos preparados. Finalmente, en la sección 6, se presentan las conclusiones del trabajo.

DESARROLLO

Breve descripción de la herramienta power query de excel

Una de las principales características de *Excel* es su flexibilidad para generar un vínculo entre datos, tablas, gráficos y demás funcionalidades, que lo hacen la hoja de cálculo más importante en los últimos tiempos.

En muchas ocasiones, al analista de datos le solicitan crear un reporte de tabla de datos en Excel y se encuentra con una gran cantidad de anomalías e inconsistencias (respecto al formato tabular) como: filas en blanco, totales y subtotales, columnas combinadas, celdas vacías, estructura no adecuada de almacenamiento de datos. Esto implica horas frente a la PC, tratando con todo tipo de fórmulas, también las matriciales, técnicas de limpieza de datos, trucos de preparación de datos y hasta la programación de macros para, finalmente, obtener un conjunto de datos de calidad que permita generar el reporte solicitado con el uso de las tablas dinámicas.

Una vía más eficiente y rápida para realizar la preparación de datos consiste en el empleo del *Power Query* (Obtener y Transformar), que es una funcionalidad nativa y totalmente embebida de Excel a partir de 2016. Esta tecnología permite preparar, adecuar, limpiar, formatear, integrar, enriquecer, crear, refinar y, hasta resumir datos de forma veloz, intuitiva y coherente (2).

Power Query está estrechamente relacionado con un proceso ETL (Extract-Transform-Load). Permite extraer datos, a partir de múltiples fuentes y organizarlos en una lista. Posteriormente, se crean transformaciones de los datos (limpiarlos, formatearlos, integrarlos y enriquecerlos) para luego cargarlos desde Power Query a una hoja Excel listos para el análisis. En resumen, Power Query convierte datos en tablas que pueden ser analizadas y además permite realizar consultas sobre ellos (2).

Ventajas de Power Query

- Permite integrar tablas de una gama de orígenes extensa y constantemente creciente de modo intuitivo y eficaz.
- Brinda bondades para combinar, anexar e incrementar tablas con pasos sencillos y acertados.
- Proporciona una interfaz intuitiva para preparar tablas, sin necesidad de pasos complejo.
- Está incorporado en las tecnologías de *Microsoft* que manipulan datos (*Microsoft Excel, Power BI de Escritorio (Desktop), Power BI en la nube (Power BI Services), Microsoft SQL Server Data Tools, Microsoft SQL Server Analysis Services, AzureAnalysis Services, Common Data Services for Apps, Microsoft flow).*

- Permite crear modificaciones y automatizaciones en el proceso ETL, sin necesidad de conocimiento avanzado.
- Power Query está en constante actualización lo que expande el potencial de la herramienta.

La figura 1 muestra una imagen del panel de opciones de *Power Query (Obtener y transformar)* incluido dentro de la pestaña Datos en *Excel*. Estas opciones permiten obtener datos de diferentes fuentes. En este caso particular, los datos se obtienen desde un libro *Excel*. También es posible obtener datos desde una tabla, un rango de celdas, permite mostrar las consultas, actualizar y ver las conexiones a las fuentes de datos.

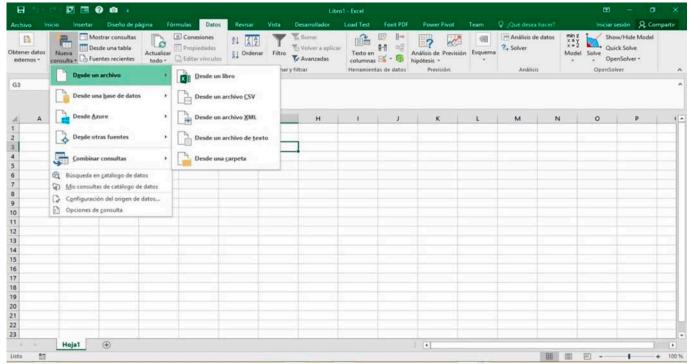


Figura 1. Imagen del panel de opciones de Power Query en Excel.

La figura 2 es una representación del editor de *Power Query* en *Excel*, con una consulta en blanco. En esta se aprecia un panel con 5 opciones.

- Archivo. Es donde se puede cerrar el editor de Power Query y cargar los resultados de la
 consulta, tanto en el destino predeterminado (nueva hoja Excel) o en el destino que se especifique. También tiene opciones para descartar los resultados de la consulta realizada y cerrar
 el editor de Power Query, así como opciones de configuración para las consultas y orígenes
 de datos.
- Inicio. Permite realizar operaciones con los datos, como separar los elementos de una columna en varias, agrupa las filas de la tabla según los valores de la columna seleccionada, cambia el tipo de dato de la columna seleccionada, promueve la primera fila de la tabla en los encabezados de las columnas, reemplaza valores existentes en las columna seleccionadas por un nuevo valor especificado, combina los resultados de diferentes consultas para obtener una nueva, anexa los resultados de una consulta debajo de los de otra.
- **Transformar.** Se pueden realizar operaciones, como eliminar duplicados, eliminar filas vacías, transponer, invertir el orden de las filas en la tabla, concatenar columnas seleccionadas en una columna, contar el número de filas de la tabla y otras.
- Agregar columna. Permite crear una nueva columna, basada en una fórmula personalizada, crear una nueva columna que agregue condicionalmente los valores en la columna seleccio-

nada actualmente, crea una nueva columna que sea el duplicado de una seleccionada, crea nueva columna basada en la concatenación de columnas seleccionadas, permite realizar operaciones estadísticas, matemáticas básicas, científicas, trigonométricas sobre las columnas.

 Vista. Se puede configurar la ventana de consulta que aparece a la izquierda en el editor de Power Query (figura 2). En esta ventana Power Query graba todos los pasos aplicados al origen de datos y se puede mover a través de ellos, para ir atrás o hacía adelante en el preprocesamiento de los datos.

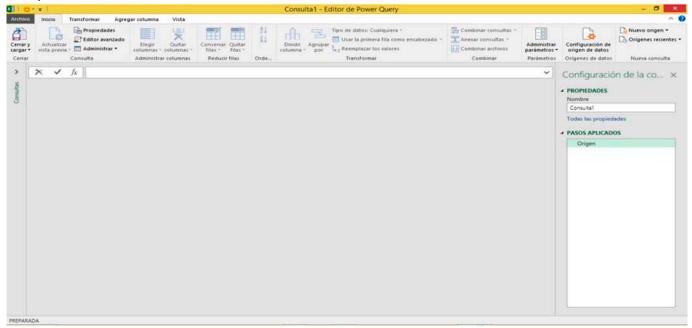


Figura 2. Imagen del editor de Power Query en Excel.

Exploración del libro de datos e identificación de inconsistencias

Se tiene información en un libro Excel sobre la cantidad de materiales que salieron del almacén, en un período de tiempo, para ser utilizados en cada área de mantenimiento industrial. Se desea construir un reporte que permita visualizar esta información.

El libro Excel entregado por el cliente contiene 12 hojas, una para cada área de mantenimiento del central azucarero: Laboratorio, Centros de acopio, Taller de maquinado, Instrumentación, Planta eléctrica, Generación de vapor, Centrífugas, Miscelánea, Basculador, Molinos y el área de Inversión de almacén.

En el rango de celdas seleccionado, de cada hoja, se ofrece la siguiente información:

- Área de mantenimiento industrial.
- Fecha de extracción del material del almacén.
- Número de norma o trabajo en el que se utilizó el material.
- Número de vale de extracción de materiales.
- Orden de trabajo asociada.
- Actividad a ejecutar.
- El encabezado de las otras columnas se corresponde con un tipo de material y la unidad de medida. En las celdas de estas columnas se escribe la cantidad de material utilizado de cada tipo.

La figura 3 muestra un ejemplo de los datos pertenecientes al área de *Centros de acopio*. Se pueden apreciar varias inconsistencias en la información que imposibilita tener datos de calidad para, posteriormente, ser analizados.

- La información está escrita en un rango de celdas y no en formato de tablas. En el formato de tablas cada columna representa una variable y, además, tiene la ventaja de que, si se adiciona una nueva fila o columna a la tabla, inmediatamente se actualiza el reporte solicitado de tabla dinámica.
- La información del Área de mantenimiento industrial está vacía y se utilizan varias columnas combinadas. Lo adecuado es emplear una sola columna para, posteriormente, hacer el reporte de tabla dinámica con la información solicitada.
- Los encabezados de las columnas están escritos en vertical, que dificulta su lectura y visualización.
- Los nombres de los materiales no están normalizados y validados. Un mismo material se puede encontrar escrito de diferentes formas.
- Los tipos de materiales empleados en cada área son diferentes por lo que el rango de celdas varía entre una hoja y otra. En total se emplearon 151 tipos distintos de materiales, por lo que se necesitaría, en cada hoja, una columna por cada material y no todos los materiales fueron utilizados en cada área de mantenimiento. Esto no permite consolidar la información de una manera eficiente.
- El formato de fecha seleccionado es *dd/mm/aaaay*; sin embargo, los valores de fecha aparecen escritos como dd.mm.aa.
- En muchas hojas aparecen materiales que no fueron utilizados.

Estas mismas inconsistencias se repiten para todas las hojas del libro, lo que hace necesario un preprocesamiento de los datos.



Figura 3. Hoja de cálculo correspondiente al área de Centros de acopio.

Pasos aplicados para el preprocesamiento de datos

Paso 1. Consistió en normalizar los nombres de las áreas de mantenimiento, los materiales y las unidades de medida. Para esto se creó una nueva hoja Excel llamada *00_Listas* que contiene 3 listas en formato de tablas (no rangos de celdas) con nombres *Lista Áreas, Lista Materiales y Lista UM*. Esto permite agregar nuevos elementos que se actualizan automáticamente en las validaciones de datos y reportes de tablas dinámicas. La figura 4 muestra una imagen de la hoja Excel.

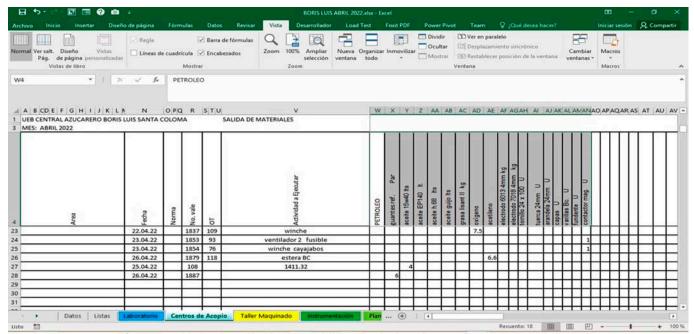


Figura 4. Hoja de cálculo con el listado de las áreas de mantenimiento, materiales y unidades de medida.

Paso 2. Se eliminaron las columnas combinadas, en cada una de las 12 hojas que tiene el libro y quedó una columna para cada encabezado. Para esto se seleccionaron las columnas combinadas y se pinchó en la opción *Combinar y central* > *Separar celdas*. Se eliminaron las columnas sobrantes y se le asignó un encabezado por cada columna. El texto de este se orientó en la dirección horizontal. Se completaron los datos de la columna Área con el nombre del área en cada fila (figura 5).

Paso 3. Se seleccionó el rango de celdas que contiene la información y se le aplicó formato de tabla, mediante la opción *Insertar>Tabla* del menú de opciones de Excel. La figura 5 muestra la imagen de la tabla creada con los datos del área de *Centros de acopio*. Este mismo procedimiento se repitió para cada una de las 12 hojas del libro Excel.

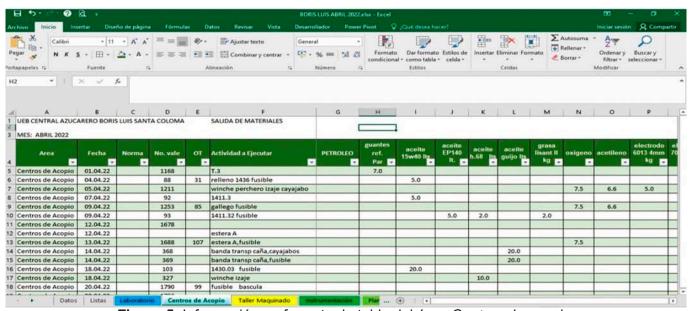


Figura 5. Información en formato de tabla del área Centros de acopio.

Paso 4. Se creó una consulta en *Power Query* por cada tabla seleccionada (Hacer click sobre la opción *Datos > Desde una tabla*). La figura 6 muestra el resultado de la consulta para la tabla

TCentros acopio. Dentro del editor de Power Query se realizaron transformaciones al formato de las columnas. Observe como los valores de la columna Fecha ya aparecen con el formato adecuado, de fecha dd/mm/aaaa. En esta tabla todavía no está resuelto el problema de dedicar una columna para cada tipo de material empleado y los tipos de materiales que se utilizan en las áreas pueden ser diferentes. Con esta estructura no se puede crear, de forma automatizada, una matriz de datos que permita consolidar, en una tabla dinámica, la información de todas las áreas.

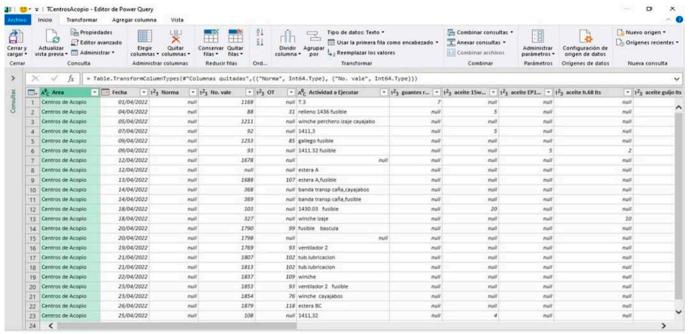


Figura 6. Información en formato de tabla del área Centros de acopio.

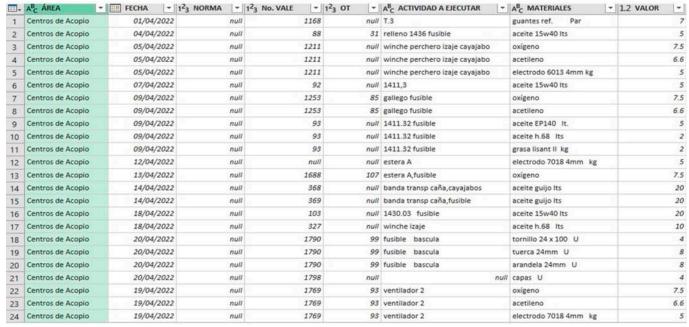


Figura 7. Información en formato de tabla del área Centros de acopio.

Paso 5. Para resolver el problema de estructura de los datos se realizó otra transformación de estos dentro del editor de *Power Query*, que permitió resumir la información de todas las columnas asociadas a los materiales en dos nuevas columnas con nombres *Materiales* y *Valor*. En la primera aparecen los nombres de los materiales y, en la segunda, la cantidad solicitada por cada vale.

Esta operación se realiza al seleccionar todas las columnas con nombres de materiales y pulsar la tecla derecha del mouse, entonces, aparece un menú emergente con la opción *Anular dinamización de las columnas seleccionadas únicamente*. La figura 7 presenta el resultado de esta transformación. Se observa que la tabla de datos del área de *Centros de acopio* tiene solamente 8 columnas.

A continuación, se procede a seleccionar la opción *Cerrar y cargar en...* y aparece el siguiente menú (figura 8) en el que el usuario indica dónde va cargar los datos. En este caso se debe indicar *Crear solo conexión*.

Este proceso se repite para todas las tablas y, finalmente, se obtienen 12 nuevas, con 8 columnas cada una.



Figura 8. Diferentes opciones para cargar los datos ya transformados en *Power Query.*

Paso 6. Ya se está en condiciones de anexar las tablas, una debajo de la otra, para crear una matriz de datos de calidad, que permita extraer la información solicitada. Esta operación se realiza de manera automática con el editor de *Power Query*, mediante la opción *Inicio > Anexar consultas>Anexar consultas para crear una nueva*. Esto permite crear una nueva consulta que contiene las tablas anexadas, de cada área, una debajo de la otra y con el mismo encabezamiento. Finalmente se pincha la opción *Cerrar y cargar* de *Power Query*, para cargar la matriz de datos en una nueva hoja Excel que se nombró *Datos* (figura 9).

UEB CENTRAL AZUCARERO BORIS LUIS SANTA COLOMA MES: ABRIL 2022											
N	AREA	FECHA	NORMA No. VAL	ORDEN TRABAJE	ACTIVIDAD A EJECUTAR	MATERIAL	VALOR	UN'	MATERIAL (UM)		
1	Laboratorio	14/04/2022	1714		sin actividad	Papel filtro	10.0	U	Papel filtro (U)		
2	Laboratorio	14/04/2022	1716		batas sanitarias	Batas sanitarias	20.0	U	Batas sanitarias (U)		
3	Laboratorio	16/04/2022	1740		sonda	Guantes latex	6.0	par	Guantes latex (par)		
4	Laboratorio	20/04/2022	1803	29	sin actividad	Varillas niquel	3.0	U	Varillas niquel (U)		
5	Laboratorio	22/04/2022	1828		sin actividad	Sacos	100.0	U	Sacos (U)		
6	Laboratorio	25/04/2022	402		prensa	Aceite hidráulico 68	4.0	It	Aceite hidráulico 68 (lt)		
7	Centros de Acopio	01/04/2022	1168		T.3	Guante reforzado	7.0	par	Guante reforzado (par)		
8	Centros de Acopio	04/04/2022	88	31	relleno 1436 fusible	Aceite 15w40	5.0	It	Aceite 15w40 (lt)		
9	Centros de Acopio	05/04/2022	1211		winche perchero izaje cayajabo	Oxígeno	7.5	m³	Oxígeno (m³)		
10	Centros de Acopio	05/04/2022	1211		winche perchero izaje cayajabo	Acetileno	6.6	m³	Acetileno (m³)		
11	Centros de Acopio	05/04/2022	1211		winche perchero izaje cayajabo	Electrodo 6013 4mm	5.0	kg	Electrodo 6013 4mm (kg)		
12	Centros de Acopio	07/04/2022	92		1411,3	Aceite 15w40	5.0	It	Aceite 15w40 (lt)		
13	Centros de Acopio	09/04/2022	1253	85	gallego fusible	Oxígeno	7.5	m³	Oxígeno (m³)		
14	Centros de Acopio	09/04/2022	1253	85	gallego fusible	Acetileno	6.6	m³	Acetileno (m³)		
15	Centros de Acopio	09/04/2022	93		1411.32 fusible	Aceite EP140	5.0	lt	Aceite EP140 (It)		
16	Contror do Aconio	09/04/2022	02		1411 22 fucible	Acoito hidráulico 69	2.0	I+	Acoito hidróulico 69 (It)		

Figura 9. Matriz final de datos en Excel.

La matriz de datos de la figura 9 tiene 389 filas y 11 columnas. Si se realiza un cambio en los datos de algunas de las tablas del libro, estos se pueden actualizar en ella al pulsar la opción Actualizar, con la tecla derecha del mouse, sin necesidad de repetir otra vez todo el proceso.

Paso 7. Se aplicó una validación de datos en la columna *Material* para hacer coincidir el nombre de los materiales con los que aparecen en la tabla *Lista Materiales*.

Paso 8. Se empleó la fórmula *BUSCARV* para completar la columna *UM*, que contiene la unidad de medida del material (Ec. 1) y concatenó la columna Material con la columna UM para obtener la columna *MATERIAL (UM)* (Ec. 2).

=BUSCARV([@MATERIAL];ListaMateriales[[MATERIALES]:[UM]];2;FALSO) Ec. 1 =[@MATERIAL]&" "&"("&[@UM]&")" Ec. 2

Extracción del conocimiento, a partir de la matriz de datos generada

A partir de la matriz de datos anterior se pulsó la opción Insertar > Tabla dinámica para generar un reporte, como el que muestra la figura 10. En el mes de abril de 2022 se consumieron 70 L de Aceite circulación 220 y fue el área de Molienda la de mayor demanda de este tipo de aceite, con 67 L en el mes. Se consumieron en el mes 113 L de Aceite hidráulico 68, de ellos, 72 L fueron entregados al área de Centrífugas. Esta tabla tiene incluida a la derecha, una barra de desplazamiento vertical, para moverse entre los diferentes materiales. Se programó una macro para actualizar automáticamente la tabla dinámica, cada vez que ocurra un cambio en la matriz de datos.

No.	áreas/material	Laboratorio	Centros de Acopio	Taller Maquinado	Instrumen tación	Planta Eléctrica	Generación de Vapor	Centrífugas	Casa Calderas	Miscelánea	Basculador	Molienda	Inversión Almacén	TOTAL
2	Aceite circulacion 220 (lt)	1.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	67.0	0.0	70.0
3	Aceite EP140 (lt)	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5.0
4	Aceite guijo (It)	0.0	40.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	38.0	0.0	78.0
5	Aceite hidráulico 68 (lt)	4.0	12.0	0.0	0.0	0.0	0.0	72.0	0.0	0.0	25.0	0.0	0.0	113.0
6	Aceite soluble (It)	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	0.0	10.0	0.0	0.0	15.0
7	Aceite turbo 68 (It)	0.0	0.0	0.0	0.0	208.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	208.0
8	Aceite viscopren (lt)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2704.0	0.0	2704.0
9	Acetileno (m³)	0.0	26.4	0.0	6.6	0.0	0.0	39.6	66.9	0.0	6.6	0.0	6.6	152.7
10	Acido (It)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1000.0	0.0	0.0	0.0	0.0	1000.0
11	Acido clorhidrico (lt)	0.0	0.0	0.0	0.0	220.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	220.0
12	Arandela 24mm (U)	0.0	8.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	8.0
13	Arandela de apresion 12mm (U)	0.0	0.0	0.0	0.0	0.0	351.0	0.0	0.0	0.0	0.0	0.0	0.0	351.0
14	Arandela plana 12mm (U)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.0	0.0	15.0
15	Arandela plana 16mm (U)	0.0	0.0	0.0	0.0	0.0	20.0	10.0	0.0	0.0	0.0	0.0	0.0	30.0
16	Arandela plana 24mm (U)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0

Figura 10. Reporte de tabla dinámica con la cantidad de materiales solicitados por área.

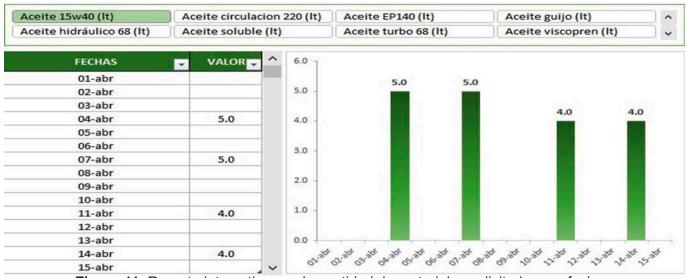


Figura 11. Reporte interactivo con la cantidad de materiales solicitados por fecha.

Otro reporte útil consiste en conocer la cantidad de materiales solicitados para trabajos de mantenimiento, por fechas. En este caso, se generó la tabla dinámica de la figura 11, que tiene asociada una segmentación de datos para seleccionar el tipo de material y tener la información por fecha. Paralelamente se construyó un gráfico dinámico de barras, para visualizar de manera interactiva la información solicitada.

CONCLUSIONES

- Se demostró la necesidad de realizar un preprocesamiento a los datos obtenidos directamente del problema.
- El análisis exploratorio de los datos permitió identificar varias inconsistencias en el libro Excel las que dificultan obtener los reportes solicitados.
- Con las transformaciones realizadas en *Power Query* se logró obtener una matriz de datos de calidad, a partir de la cual se obtienen los reportes de resultados.
- Se construyó un reporte de tabla dinámica que brinda información sobre la cantidad de materiales de cada tipo, solicitado por área, en el período analizado, así como el total general.
- El reporte interactivo permite consolidar la información por fecha de solicitud de materiales.

REFERENCIAS BIBLIOGRÁFICAS

- 1. García, S., Ramírez, S., Luengo, J., Herrera, F. *Big Data:* Preprocesamiento y calidad de datos. Big Data monografía., julio-octubre 2016, España, ATI, Novática No. 237, pp.17-23. ISSN: 2444-6629.
- 2. Caballero, M., Torres, F.EL ADN de Power Query. Power Bi Aplicado a Limpieza de Datos Inteligente. Primera Edición COLECCIÓN ADN, Colombia 2019. ISBN: 978-958-46-8524-7, pp. 638.