

# Data Wrangling: conceptos básicos y operaciones

---

*El objetivo de esta lectura es comprender en qué consiste el **Data Wrangling** y entender algunas de las técnicas más empleadas.*

---



## ¿En qué consiste el *Data Wrangling*?

El **Data Wrangling** es un proceso que consiste en la correcta preparación (almacenamiento y transformación) de los datos en bruto, con el fin de asegurar su calidad y disponerlos otro formato más accesible para su posterior análisis (u otro uso que se les quiera dar).

El grado de calidad y accesibilidad conseguido para los datos mediante este proceso será determinante para que su análisis permita obtener resultados útiles y coherentes.

Por ejemplo, imagina que eres propietario de un negocio de fabricación de chocolate. Comenzaste tu actividad con una antigua fábrica de chocolate y, en los últimos diez años, has hecho prosperar tu negocio abriendo cinco fábricas más. A medida que tu negocio crecía, has ido añadiendo nuevos sistemas de información en cada fábrica para controlar y gestionar tu producción, y almacenando toda esa información en tus repositorios de información.

Como parte de tu estrategia de negocio, te planteas utilizar las tecnologías Big Data para analizar los últimos dos años de datos de producción de tu compañía y encontrar formas de reducir los costes de operación.

En tus bases de datos tienes mucha información para plantear el análisis, pero dado que cada una de tus cinco fábricas opera en un sistema de información diferente, la información se almacena en repositorios diferentes, no siempre es la misma para el mismo proceso, y en muchas ocasiones tiene formatos diferentes. Además, para unas fábricas tienes datos almacenados con una frecuencia diaria, y en otras, con una frecuencia semanal.

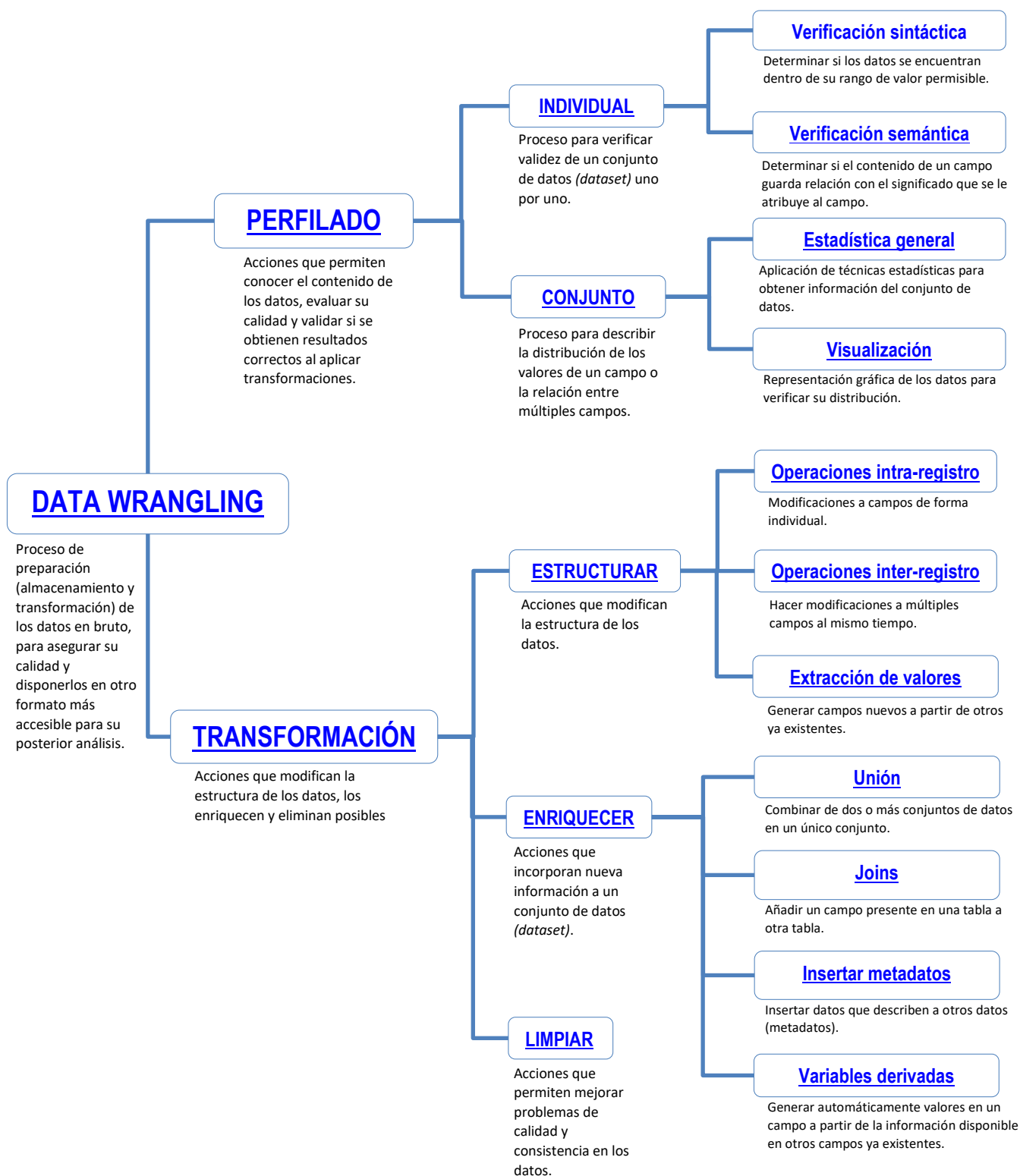
Ser capaz de trabajar estos datos y sacar conclusiones a partir de ellos para tomar mejores decisiones en tu estrategia de negocio puede suponer un auténtico reto debido a la heterogeneidad de los datos.

Los procesos de **Data Wrangling** te permitirán preparar los datos y colocarlos en formato más accesible y, así, podrías llegar a tener fuentes de datos mucho más limpias y homogéneas. Por ejemplo, todos los datos de producción con el mismo formato, expresados en los mismos lapsos de tiempos, entre otros y comenzar a analizarlas con mayor sencillez.

Para facilitar tu comprensión y navegación de esta lectura, hemos incluido un mapa conceptual en la siguiente página.

# Mapa de la lectura

NOTA: haz clic sobre los conceptos en el mapa para navegar la lectura. Puedes volver a este mapa cuando quieras, haciendo clic sobre el texto “Ir a ‘Mapa de la lectura’” ubicado al final de cada página del documento.



## Acciones básicas del *Data Wrangling*

A continuación, se definen las dos acciones básicas del ***Data Wrangling***: 1) perfilado y 2) transformación.

### 1. Perfilado

El perfilado es un conjunto de acciones que permiten conocer el contenido de los datos, evaluar su calidad y validar si se obtienen resultados correctos al aplicar transformaciones.

Este proceso es importante, porque los datos que maneja tu organización pueden tener errores que, si no son resueltos, luego se transmiten a cualquier análisis que quieras hacer con los datos, o lo limitan en mayor o menor medida. Es importante reportar los resultados de los procesos de perfilado, no solo para conocer los datos que tienen errores y resolverlos, sino para prevenirlos. Por ejemplo, si tenemos una lista de e-mails que provienen de un formulario rellenado por usuarios y encontramos que muchos e-mails están errados porque usuario no incluyó el símbolo arroba (@) o un dominio de correo electrónico válido (gmail.com, yahoo.com, etc.), entonces conviene que en el formulario sea obligatorio que el texto sea una dirección de correo válida porque contiene tanto el símbolo @ como un dominio de correo electrónico válido.

Existen dos tipos de perfilado: el perfilado individual y el perfilado de conjunto. A continuación, te los explicamos con ejemplos.

1.1. Perfilado individual: proceso que consiste en *verificar la validez de un conjunto de datos (dataset)* uno por uno, dato por dato. Existen dos tipos de técnicas que se usan para hacer el perfilado individual: la *validación sintáctica* y la *validación semántica*. A continuación, te las definimos con ejemplos:

- Verificación sintáctica: se emplea para determinar si los datos contenidos en un campo son válidos dentro de las restricciones de contenido que se han definido para ese campo. Dicho de otra forma, consiste en verificar si los datos se encuentran dentro de su rango de valor permisible.

Por ejemplo, si en las bases de datos de los clientes de tus fábricas de chocolate se ha definido un campo que exprese si el cliente es mayor de edad, y este campo sólo admite un 1 que signifique la mayoría de edad, o un 0 cuando el cliente sea menor; cualquier valor diferente de 1 o 0 será un error (esta característica de un campo de datos, que permite sólo dos valores, es lo que lo define como campo booleano).

Edad	Mayoría de edad	Resultado de la verificación sintáctica
18	1	Dato válido
30	1	Dato válido
10	0	Dato válido
9	0	Dato válido
21	2	Dato inválido

Esta identificación de errores es importante porque, de no ser corregidos, el análisis que quieras hacer posteriormente de estos datos mostrará conclusiones erróneas, o incluso podría derivar en impedir el análisis de los datos inválidos. Si no corrigieras ese “2” por un “1”, puede que ese cliente de 21 años no sea tomado en cuenta por otras herramientas de análisis en el futuro y tu análisis estaría errado.

- o Verificación semántica: se emplea para determinar si el contenido de un campo guarda relación con el significado que se le atribuye al campo. Por ejemplo, un campo definido “Ciudad” debería contener solo valores que semánticamente estén relacionados con ciudades, es decir, valores como “Madrid” o “Barcelona”; pero no podría contener valores como “Argentina”, porque corresponde con un país y no una ciudad, o “brasileño”, que constituye una nacionalidad y no una ciudad.

Ciudad	Resultado verificación semántica
Madrid	Dato válido
Brasileño	Dato inválido
Buenos Aires	Dato válido
Argentina	Dato inválido

Por ejemplo, si tienes clientes que viven en distintos países, vas a necesitar que los campos de sus direcciones estén correctos (país, ciudad, dirección, etc.). La verificación semántica te ayudará a detectar dónde hay datos incorrectos.

1.2. Perfilado de conjunto: es un proceso que consiste en describir la distribución de los valores de un campo, con el fin de identificar valores atípicos que puedan señalar potenciales incorrecciones en los datos. En otras palabras, consiste en determinar cómo se reparten los valores de un determinado campo en el rango que se ha definido para este campo. Por ejemplo, para un campo “edad del cliente”, el perfilado de conjunto permite detectar si la mayoría de los datos se concentran en el rango entre los 20 y 25 años mientras que algunos valores se sitúan entre los 30 y 40 años. De este modo, un valor de edad de 150 años, atípico en esa distribución de valores, puede identificarse como un error en el registro de los datos.

Existen dos tipos de técnicas que se usan para hacer el perfilado de conjunto: la *estadística general* y la *visualización*. A continuación, te las definimos con ejemplos:

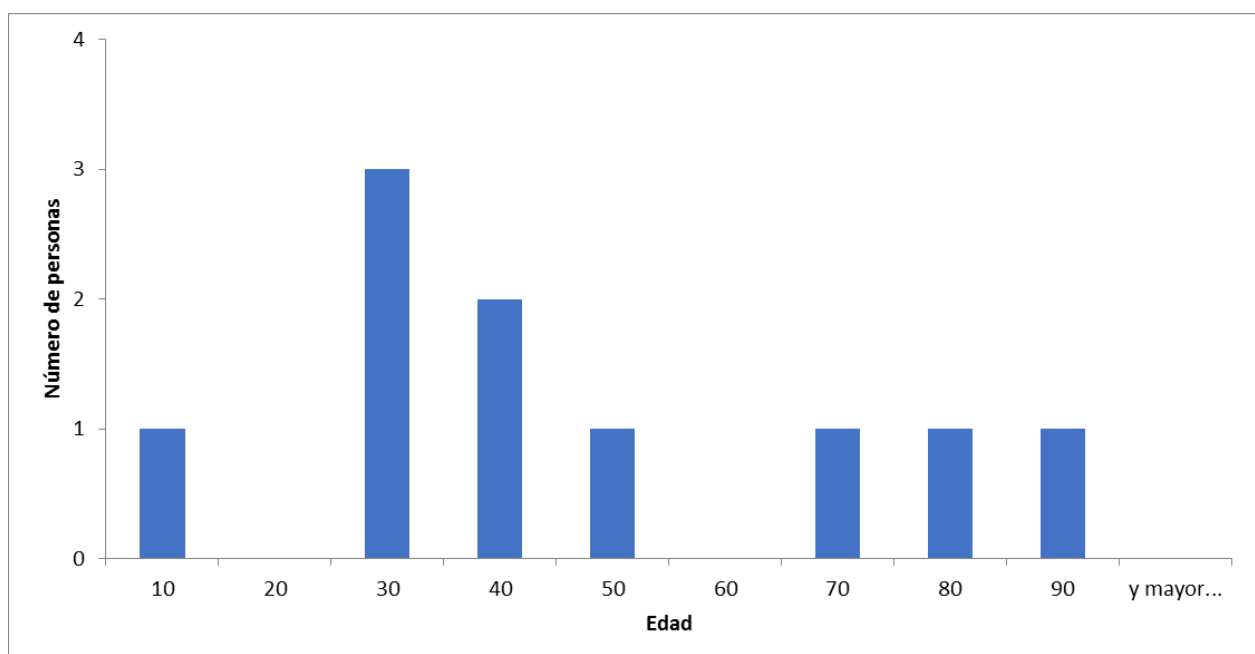
- o Estadística general: consiste en la aplicación de técnicas estadísticas para obtener información del conjunto de datos, tales como la media, valores máximos y mínimos, desviación estándar, etc. Por ejemplo, si se consideran los datos de edad que se comentaban anteriormente, el perfilado estadístico arroja resultados como los de la tabla:

Datos edad: 21, 23, 35, 21, 39, 7, 45, 61, 72, 83

Parámetro estadístico	Valor (edad)
Media aritmética (promedio)	40,7
Mediana	37
Valor máximo	83
Valor mínimo	7
Desviación típica	24,63

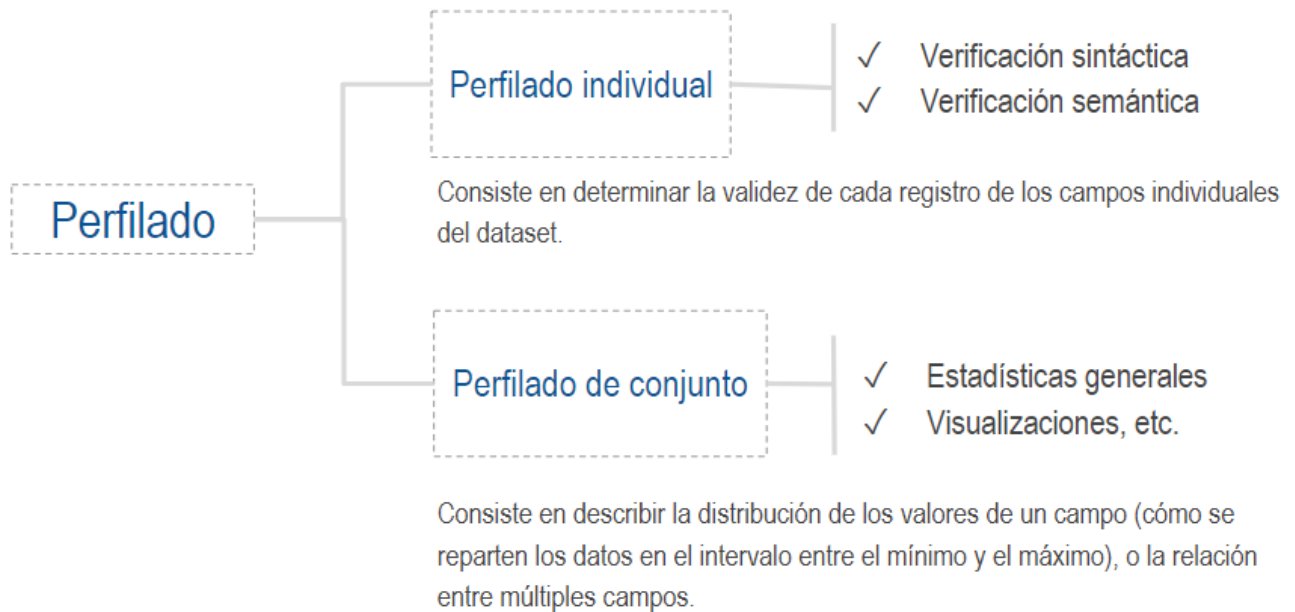
Tener la edad promedio de tus clientes consumidores de chocolate te puede ayudar a orientar tu producción y negocio para satisfacerles de mejor manera, por ejemplo. En ese sentido, la estadística general puede ser una herramienta muy útil.

- Visualización: consiste en la representación gráfica de los datos para verificar la distribución de los datos (por ejemplo, si existe alguna tendencia). En el caso de datos numéricos, permite compararlos con patrones estadísticos conocidos. Con datos como las edades de tus clientes de tu fábrica más antigua, por ejemplo, la representación gráfica permite identificar rápidamente que el mayor grupo de personas en un mismo rango de edad es el que corresponde a jóvenes entre 20 y 30 años, que existen personas en el rango entre los 30 y 40 años, y que no hay personas mayores de 90 años.



Las visualizaciones pueden ser útiles para detectar rápidamente patrones o tendencias que, de otra forma, son más difíciles de detectar, y posibles irregularidades en los datos. Si en vez de un gráfico tuvieras una tabla con miles de datos correspondientes a las edades de tus clientes, sería mucho más difícil detectar en la tabla que la mayoría de tus clientes tienen entre 30 y 50 años.

A continuación, te presentamos una tabla con una síntesis de las acciones de “perfilado” correspondientes a **Data Wrangling**:





## 2. Transformación

La transformación es la segunda acción básica del **Data Wrangling**. Consiste en el conjunto de acciones que modifican la estructura de los datos, enriquecen los datos y eliminan errores de los datos.

Para generar la transformación, se pueden ejecutar tres grandes acciones: estructurar, enriquecer y limpiar. A continuación, te explicaremos una por una, con ejemplos, y las técnicas a utilizar en cada una:

2.1 Estructurar: es cualquier acción que modifica la estructura o *schema* de los datos. La estructura de los datos hace referencia a cómo se organizan y almacenan en la base de datos (por ejemplo, en filas y columnas formando una tabla con títulos). Esta acción considera tres tipos de técnicas: operaciones intra-registro, operaciones inter-registro y extracción de valores. Veamos cada una de ellas y sus respectivos subtipos:

- 1) *Operaciones intra-registro*: se emplean para realizar *modificaciones sobre campos de forma individual*. Por ejemplo, una operación intra-registro sería reordenar de una lista de datos haciendo que el que está en último lugar pase al principio, o crear un campo calculado como la suma de otros dos.

Por ejemplo:

*Dataset* inicial:

2	4	5,4	0	7
---	---	-----	---	---

*Dataset* tras reordenación de valores (el último valor “7” pasa al principio de la tabla y los demás valores se mueven de lugar):

7	2	4	5,4	0
---	---	---	-----	---

El ejemplo anterior, aplicado a nuestra fábrica de chocolates, sería útil para asegurar que la información de la producción está ordenada siguiendo un determinado orden.

- 2) *Operaciones inter-registro*: consiste en realizar *modificaciones sobre múltiples campos al mismo tiempo*. Por ejemplo, eliminar los datos que tengan una antigüedad mayor de dos años, o definir los valores del precio de la vivienda a nivel de comunidad autónoma a partir de los valores del precio de la vivienda a nivel de municipio.

Por ejemplo:

*Dataset inicial:*

Área geográfica	Número de habitantes
Madrid	3,166 M
Londres	8,788 M
Buenos Aires	2,891 M
México D.F.	8,851 M

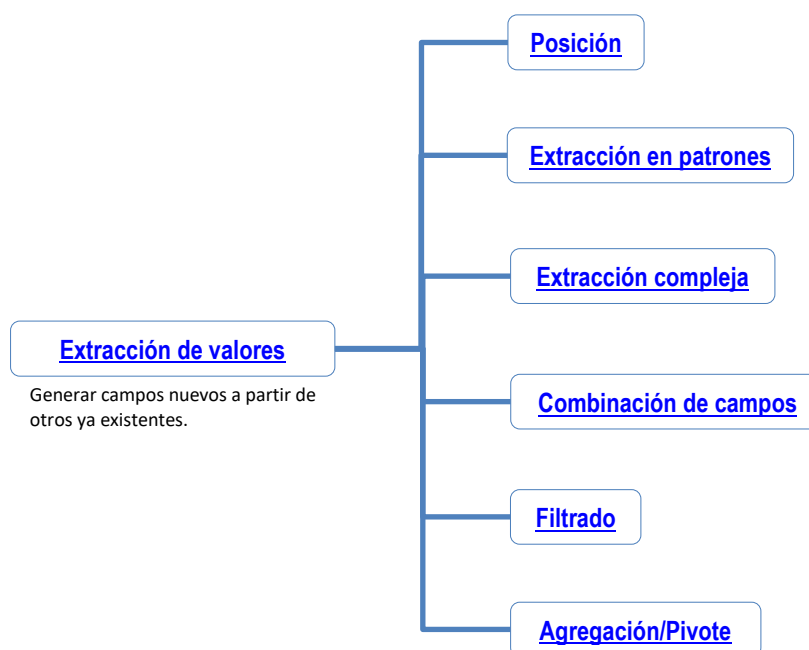
*Dataset tras operación interregistro (se agregan datos por continente):*

Área geográfica	Número de habitantes
Europa	11,954 M
América	11,742 M

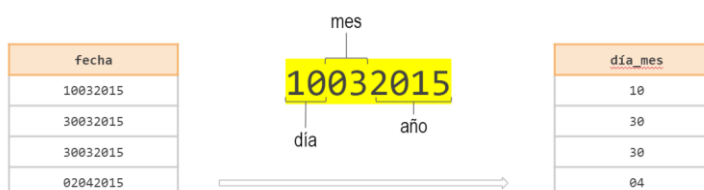
El ejemplo anterior, aplicado a nuestra fábrica de chocolates, sería útil para obtener cifras de nuestras ventas agregadas a nivel de continente y poder disponer de esta información para análisis posteriores.

- 3) **Extracción de valores**: se emplean para *extraer datos a partir de cadenas de información mayores*, de tal forma que se generan campos nuevos a partir de campos ya existentes.

A continuación, te presentamos un mapa conceptual con todos los criterios que pueden seguirse para “Extracción de valores” (ver siguiente página):



1. Posición: por ejemplo, si un campo contiene la concatenación del día, mes y año como una secuencia de números (por ejemplo, 10032015 para 10/03/2015), una extracción por posición para obtener el día se basa en recoger los dos primeros dígitos de la secuencia:



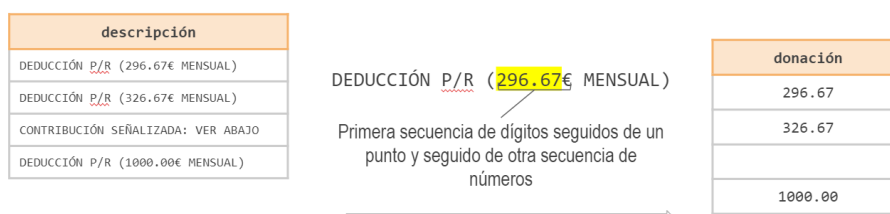
El ejemplo anterior, aplicado a nuestra fábrica de chocolates, sería útil para desagregar los datos de fecha que registra una máquina en su log de funcionamiento, para poder explotarla en un análisis posterior.

[Volver a mapa "Extracción de valores"](#)

2. Extracción basada en patrones: por ejemplo, de una línea de texto donde se indica un importe seguido del símbolo del euro, el patrón para extraer el valor del importe consiste en buscar los dígitos de la

cadena que se corresponden con números antecediendo al símbolo del euro, que tienen el símbolo decimal.

Por ejemplo, del campo texto DEDUCCIÓN P/R (296.67€ MENSUAL) se extrae el valor 296.67 para registrarlo en un campo importe.



El ejemplo anterior, aplicado a nuestra fábrica de chocolates, sería útil para registrar en nuestra base de datos el importe de las deducciones en una factura de nuestros proveedores de materia prima para la elaboración de chocolate.

[Volver a mapa "Extracción de valores"](#)

3. Extracción compleja: se corresponde con el caso en el que los datos a extraer no están recogidos en una única cadena, como es el caso de datos semiestructurados, donde la estructura de formato, tamaño y longitud no es fija. Por ejemplo, si en una factura electrónica se han definido como metadato (datos que describen otro dato) el proveedor, el tamaño y el periodo facturado:

Concepto factura: "Provisión de chocolate en polvo"

Proveedor: "Chocolates S.A."

Periodo facturado: 01/11/2015 - 01/12/2015

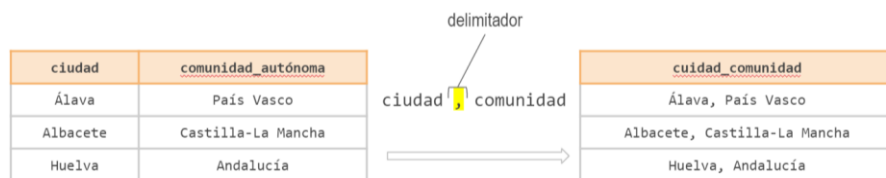
Tamaño: 450 MB

Una extracción compleja puede ser obtener el periodo en el que se ha emitido la factura, "01/11/2015 - 01/12/2015", para por ejemplo

poder indicar en el sistema de contabilidad que se ha recibido esta factura.

[Volver a mapa “Extracción de valores”](#)

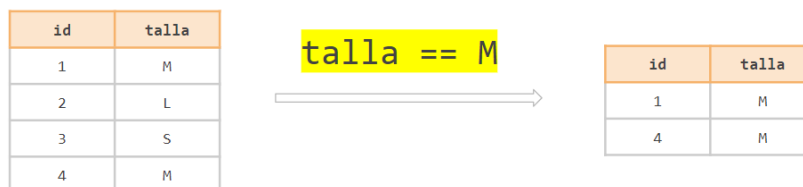
- Combinación de campos: por ejemplo, formar un único campo “Ciudad\_Comunidad” para denotar la dirección a partir de los campos “Ciudad” y “Comunidad”:



El ejemplo anterior, aplicado a nuestra fábrica de chocolates, puede ser útil para mostrar en un informe más detalle sobre la localización de nuestros clientes.

[Volver a mapa “Extracción de valores”](#)

- Filtrado: se emplea para establecer criterios que permitan la acción de otras acciones, por ejemplo, eliminación en masa de datos o selecciones en base a una pauta para cambiar el nivel de detalle de un conjunto de datos (granularidad).



El filtrado es útil cuando se desea estudiar únicamente información que cumple unos criterios definidos, por ejemplo, los clientes de nuestra fábrica de chocolate que están entre 20 y 30 años.

[Volver a mapa “Extracción de valores”](#)

6. Agregación/Pivote: se emplea para modificar el nivel de detalle (granularidad) de un conjunto de datos (como, por ejemplo, sumar todos los importes de una lista de costes para obtener el coste total).

Por ejemplo, mira el siguiente set de datos referentes a las compras de una compañía a proveedores:

Trimestre	Proveedor 1	Proveedor 2	Proveedor 3
Primer Trimestre	10.652,10 €	12.782,52 €	1.775,35 €
Segundo Trimestre	13.847,73 €	16.617,28 €	2.307,96 €
Tercer Trimestre	2.130,42 €	2.556,50 €	355,07 €

Se realizan dos operaciones: una agregación de totales y un pivote fila-columna:

Trimestre	Primer Trimestre	Segundo Trimestre	Tercer Trimestre
Proveedor 1	10.652,10 €	13.847,73 €	2.130,42 €
Proveedor 2	12.782,52 €	16.617,28 €	2.556,50 €
Proveedor 3	1.775,35 €	2.307,96 €	355,07 €
Total	25.209,97 €	32.772,96 €	5.041,99 €

Las acciones del ejemplo anterior, aplicado a nuestra fábrica de chocolates, resultaría de gran utilidad para saber cuáles han sido los costes totales por compras a proveedores en cada uno de los trimestres, o poder visualizar en un informe de nuestro departamento financiero a la dirección de las fábricas la información enfocada al proveedor o al trimestre, según se desee.

[Volver a mapa "Extracción de valores"](#)

- 2.1 Enriquecer: corresponde a las acciones que permite incorporar nueva información al *dataset*, con el fin de facilitar el análisis posterior de los datos. Puede consistir en insertar registros o campos desde otros *datasets* relacionados o usar fórmulas para crear nuevos campos.

- 1) *Unión*: consiste en la *combinación de dos o más conjuntos de datos en un único conjunto*, mediante la adición de registros adicionales en un conjunto de datos ya existente. Por ejemplo, si se tienen registrados los ingresos por fecha de una pequeña tienda de barrio y se ha guardado la información en dos tablas diferentes, puede realizarse una unión para conseguir una única tabla con todos los ingresos.

fecha	ingresos
01032015	34
02032015	77
...	...

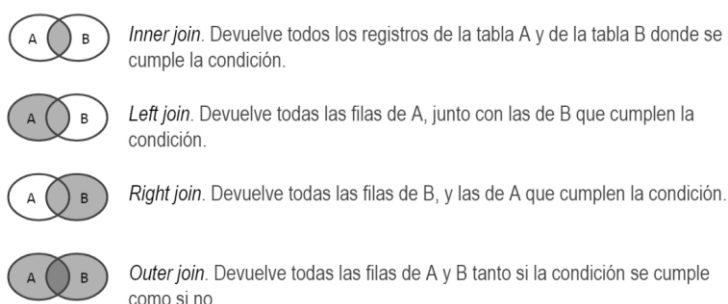
fecha	ingresos
01042015	150
02042015	23
...	...

fecha	ingresos
01032015	34
02032015	77
...	...
01042015	150
02042015	23
...	...

En nuestra fábrica de chocolates, esta operación sería útil para guardar en una única tabla la información de ingresos para cada una de las fábricas de nuestro grupo.

- 2) *Joins*: en el caso de que el conjunto de datos sean *tablas*, es posible *enriquecer una de ellas añadiendo una fila o columna a partir del conjunto de datos* que resulta de combinar total o parcialmente estas tablas. Esta unión se realiza en base a un campo común.



En el gráfico anterior, A y B denotan tablas de datos, y la tabla resultado del *join* o unión es la que está representada por el área sombreada.

Por ejemplo, si en una tabla están disponibles los datos personales de un alumno, y en otra tabla las notas obtenidas de los alumnos, puede añadirse a la primera las notas del alumno a través de un *join* de estas dos tablas, donde el campo común es el identificador del alumno.

Identificador alumno	Nombre	Edad
12387425	Juan Antonio Martínez	21
63978510	María Clara Ramírez	25
54257441	Rosa González	19

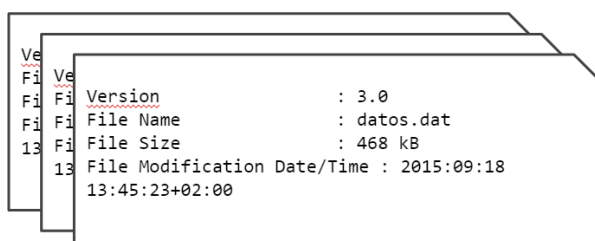
Identificador alumno	Nota media curso
12387425	8.5
63978510	7.3
54257441	9.4

Tras el *join*, se obtiene el siguiente *dataset*:

Identificador alumno	Nombre	Edad	Nota media curso
12387425	Juan Antonio Martínez	21	8.5
63978510	María Clara Ramírez	25	7.3
54257441	Rosa González	19	9.4

Como se observa, esta operación es útil para mantener en una sola tabla información relevante para un análisis posterior, facilitándolo en gran medida.

- 3) *Insertar metadatos*: la *inserción de metadatos* (datos que describen otros datos) resulta de gran utilidad para añadir significado a un dato o favorecer su accesibilidad. Los más comunes suelen ser el nombre del fichero, número de registros, y fecha y hora de creación/actualización/acceso.



Version	: 3.0
File Name	: datos.dat
File Size	: 468 kB
File Modification Date/Time	: 2015:09:18 13:45:23+02:00



- 4) *Variables derivadas*: consiste en *generar valores en un campo a partir de la información disponible en otros campos*, por ejemplo, la estación a partir de la fecha. En este ejemplo, a la fecha 04/06/2018 le corresponde otoño (hemisferio sur).

Fecha venta	Estación del año
04/06/2018	Otoño
12/012/2015	Verano

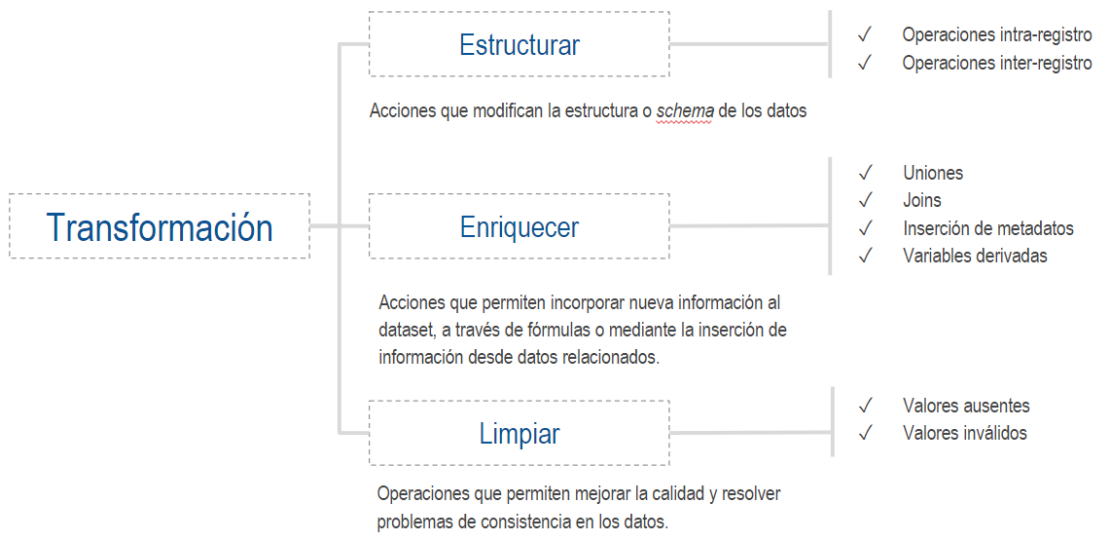
Esta operación, en el ejemplo de nuestra fábrica de chocolates, resulta de gran utilidad para estudiar la estacionalidad de nuestras ventas.

2.2 Limpiar: son operaciones que permiten *mejorar la calidad y resolver problemas de consistencia en los datos*. Consiste en tareas de manipulación de los registros para, entre otras cosas, tratar valores nulos o inválidos.

En relación a la limpieza de datos, las acciones habituales tienen que ver con depurar los datos de mala calidad, bien sea señalando o eliminando valores inválidos detectados (por ejemplo, inconsistencias detectadas durante el perfilado), valores nulos que no aportan información relevante, información duplicada, etc.

En el ejemplo de nuestra fábrica de chocolates, las labores de limpieza pueden eliminar, por ejemplo, los registros de fechas que no cumplen el formato de fecha establecido, o los valores inválidos del ejemplo de mayoría de edad que se indicó en el perfilado.

A continuación, te presentamos una tabla con una síntesis de las acciones de “transformación” correspondientes a **Data Wrangling**:



## Conclusiones

Corresponde al proceso del **Data Wrangling** el almacenamiento de los datos, su procesamiento y la preparación de los mismos. Este proceso resulta realmente importante en proyectos Big Data ya que permite optimizar el proceso de limpieza y almacenamiento de datos para que queden preparados con la máxima calidad posible para el análisis.



Esta obra está sujeta a la Licencia Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-nd/3.0/es/> o envíe una carta Creative Commons, PO Box 1866, Mountain View, CA 94042, USA.