

BIG DATA Business

2.7 Data Wrangling

Conecta Empleo

Contenido desarrollado por
Synergic Partners



An abstract network diagram composed of teal-colored nodes and connecting lines. The nodes are scattered across the left side of the slide, with some forming small, dense clusters and others existing as isolated points or simple line segments. The lines connect the nodes in a non-uniform, organic pattern.

Definición

CONCEPTOS BÁSICOS DE DATA WRANGLING

Definición

“

Data Wrangling proceso de unificación y limpieza de datos para facilitar su acceso y posterior análisis. Se enmarca dentro de la fase de preprocesado. Dados los datos en crudo, se realiza un proceso de conversión o mapeado a otro formato que permita acceder a los datos de forma más organizada. Tras identificar los datos que se quieren explorar, será necesario aplicar técnicas de extracción y transformación para un buen tratamiento en las siguientes fases del *pipeline*.

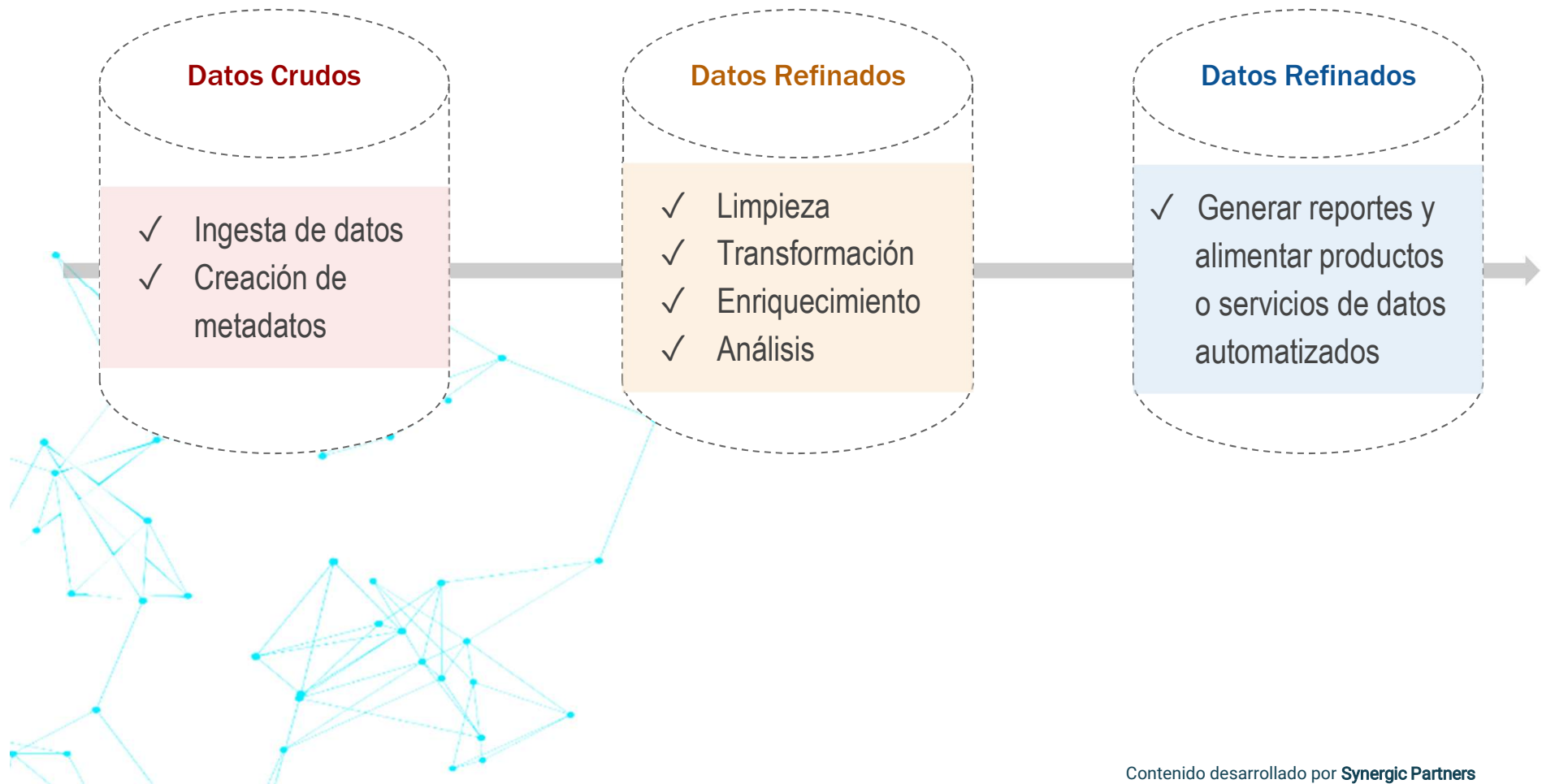
CONCEPTOS BÁSICOS DE DATA WRANGLING

Importancia



CONCEPTOS BÁSICOS DE DATA WRANGLING

Flujo del Dato



An abstract network diagram consisting of numerous teal-colored nodes connected by thin teal lines. The nodes are scattered across the left side of the slide, with some forming small, dense clusters and others existing as isolated points or part of larger, more complex web-like structures. The lines vary in length and orientation, creating a sense of dynamic connectivity.

Etapas de Data Wrangling

CONCEPTOS BÁSICOS DE DATA WRANGLING

Etapas de Data Wrangling

1

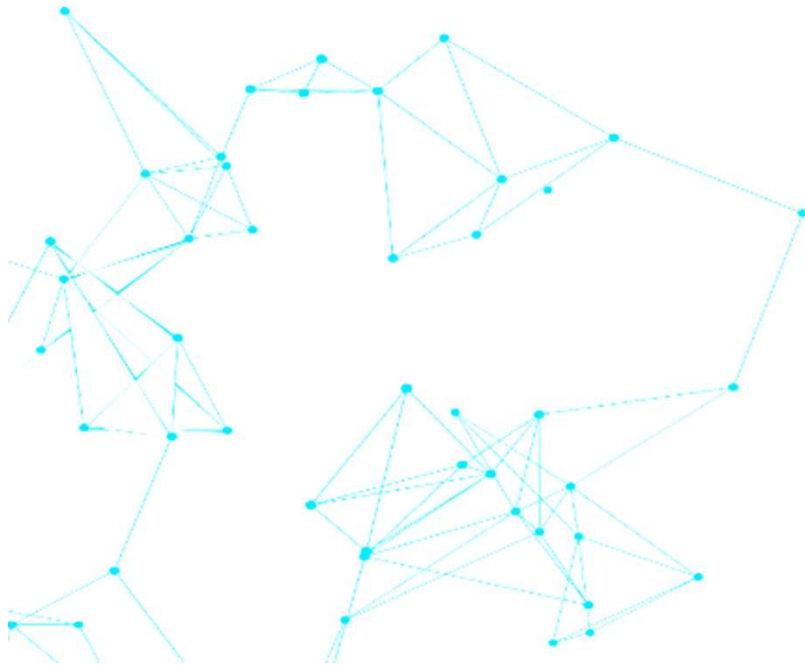
Acceso a los datos

2

Transformación

3

Publicación



CONCEPTOS BÁSICOS DE DATA WRANGLING

Etapas de Data Wrangling

1

Acceso a los datos

- ✓ Obtención de permisos de acceso
- ✓ Cambios en la infraestructura de los datos
- ✓ Manipulación de la localización y relaciones entre datasets

2

Transformación

- ✓ Acciones de manipulación de la estructura, granularidad, temporalidad, y alcance de los datos

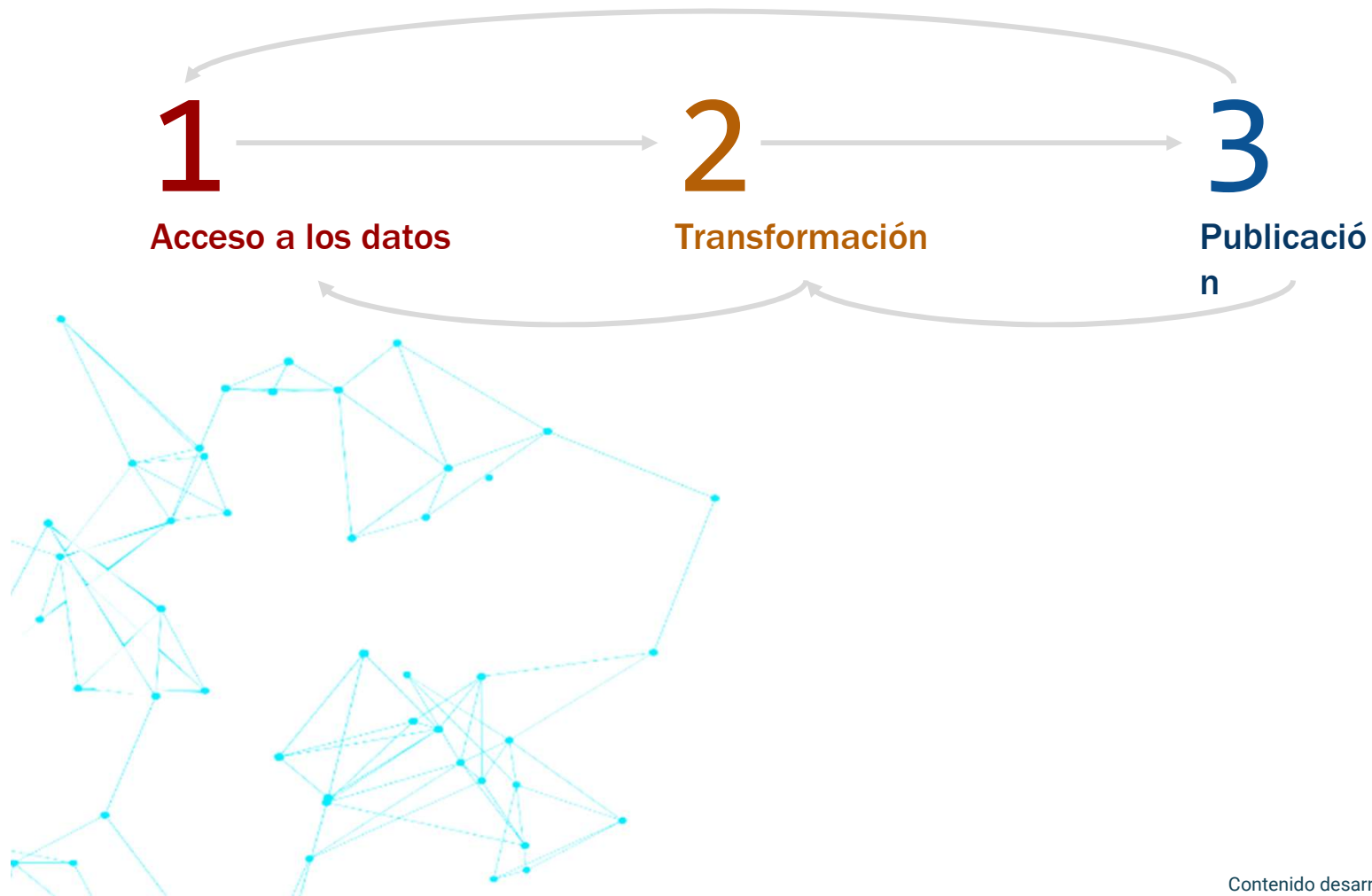
3

Publicación

- ✓ Datasets transformados
- ✓ Scripts con lógica de transformación
- ✓ Metadata descriptiva del dataset

CONCEPTOS BÁSICOS DE DATA WRANGLING

Etapas de Data Wrangling



An abstract network diagram composed of teal-colored nodes (small dots) and lines (edges) connecting them. The nodes are scattered across the left side of the slide, with some forming small clusters and others being isolated. The lines vary in length and orientation, creating a complex web-like structure.

Acciones básicas de Data Wrangling

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas de Data Wrangling

Perfilado

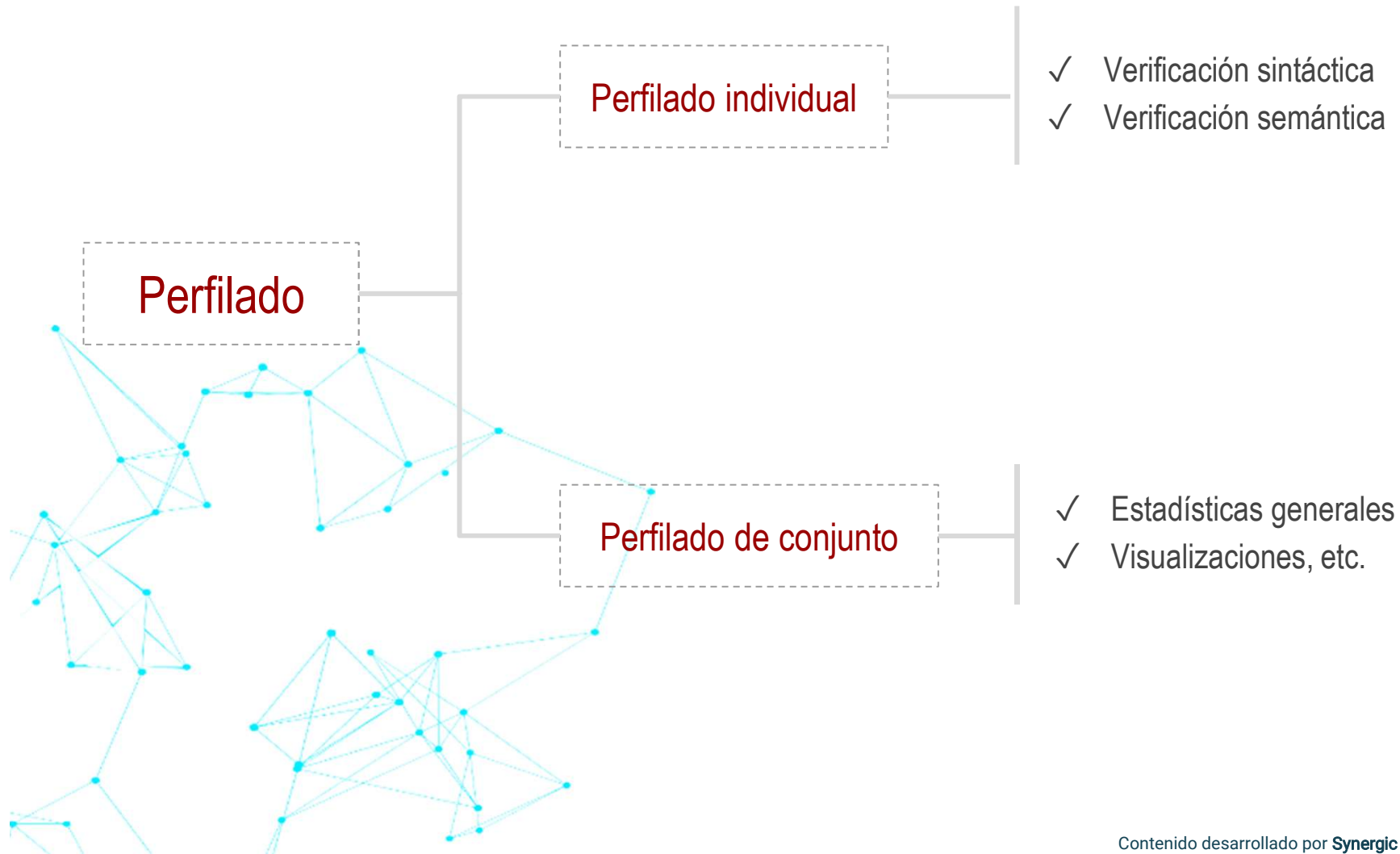
Conjunto de acciones que permiten conocer el contenido de los datos, evaluar su calidad y validar si se obtienen resultados correctos al aplicar transformaciones.

Transformación

Conjunto de acciones que modifican la estructura, enriquecen o eliminan errores de los datos.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas de Data Wrangling



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Perfilado

Perfilado individual

Consiste en determinar la validez de cada registro de los campos individuales del dataset.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Perfilado

Perfilado individual

Verificación sintáctica. La sintaxis se refiere a los valores literales o el rango de valores válidos de un campo. El perfilado para restricciones sintácticas consiste en verificar si los datos se encuentran dentro su rango de valor permisible.

Ejemplos:



Valores booleanos codificados como bits
{0, 1}



Sexo {Hombre, Mujer}



Total minutos llamadas diarias

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Perfilado

Perfilado individual

Verificación semántica. Las restricciones de tipo semántico se refiere al significado o interpretación de los valores de un campo. En general, requiere derivar un nuevo campo que codifique explícitamente la interpretación semántica del campo fuente.

Ejemplos:



El campo **Ciudad** contiene valores:

- Madrid
- MAD
- Madrd



El campo **Edad** contiene valores -1 para indicar que un valor ausente.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Perfilado

Perfilado en conjunto

Consiste en describir la distribución de los valores de un campo, o la relación entre múltiples campos.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Perfilado

Perfilado en conjunto

Ejemplos:



Para campos numéricos la distribución de los datos se puede verificar a partir de un histograma y comparándolo con distribuciones estadísticas conocidas.



Estadísticas generales como: valores mínimos y máximos, media, mediana, desviación estándar.



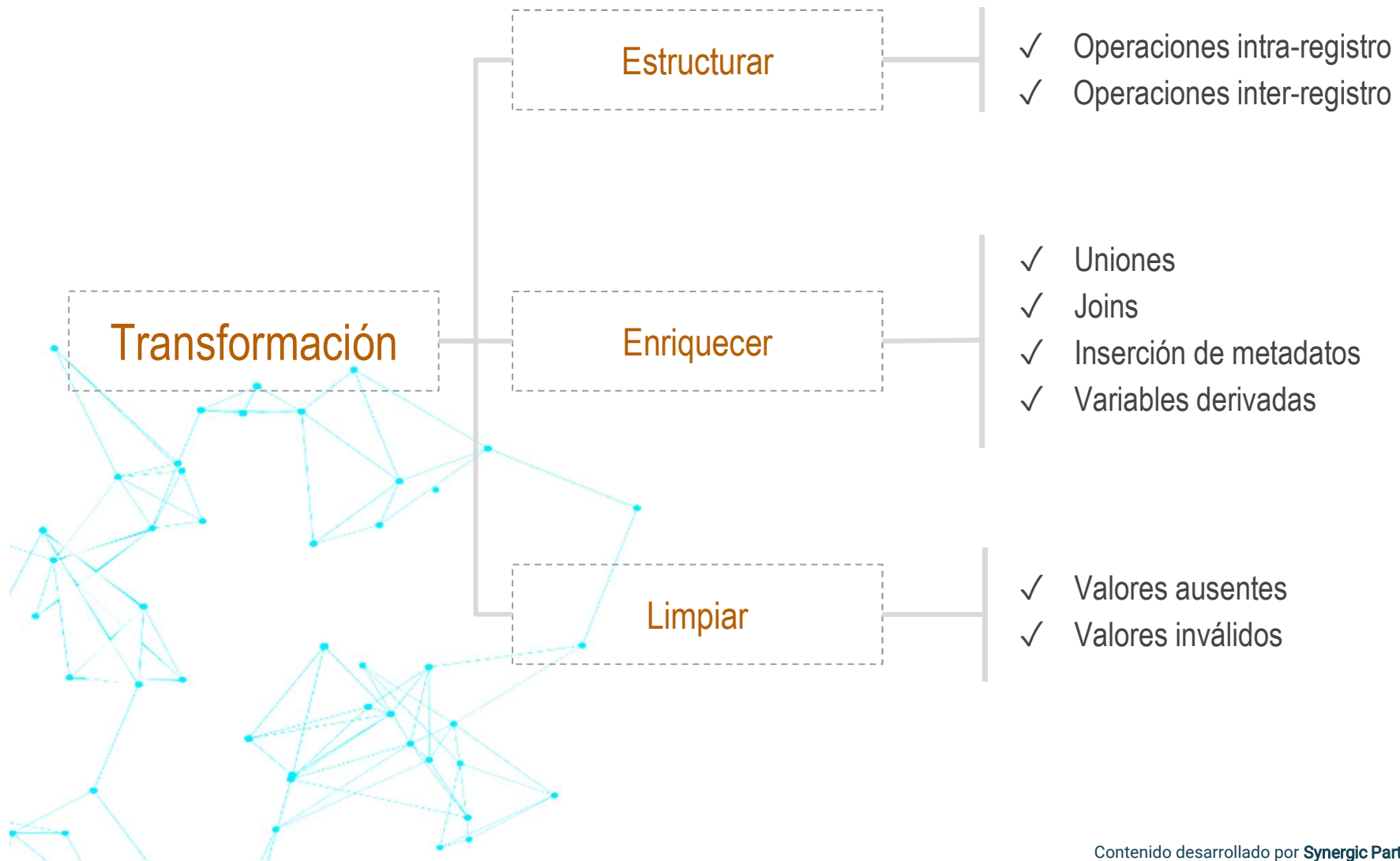
Representación en mapas de datos geoespaciales



Número ocurrencias de valores únicos.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar

Consiste en cualquier acción que modifica la forma o *schema* de los datos. Las acciones de este tipo pueden ser:

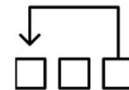


CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar

Operaciones intra-registro. Conjunto de operaciones de manipulación de nivel de registros y campos individuales.



Reordenación de campos.



Creación de nuevos campos a partir de extracción de valores.



Combinación de múltiples campos en un solo campo.

CONCEPTOS BÁSICOS DE DATA WRANGLING

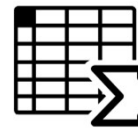
Acciones básicas. Transformación

Estructurar

Operaciones inter-registro. Conjunto de operaciones de manipulación sobre múltiples campos a la vez.



Eliminación de registros mediante filtrado.



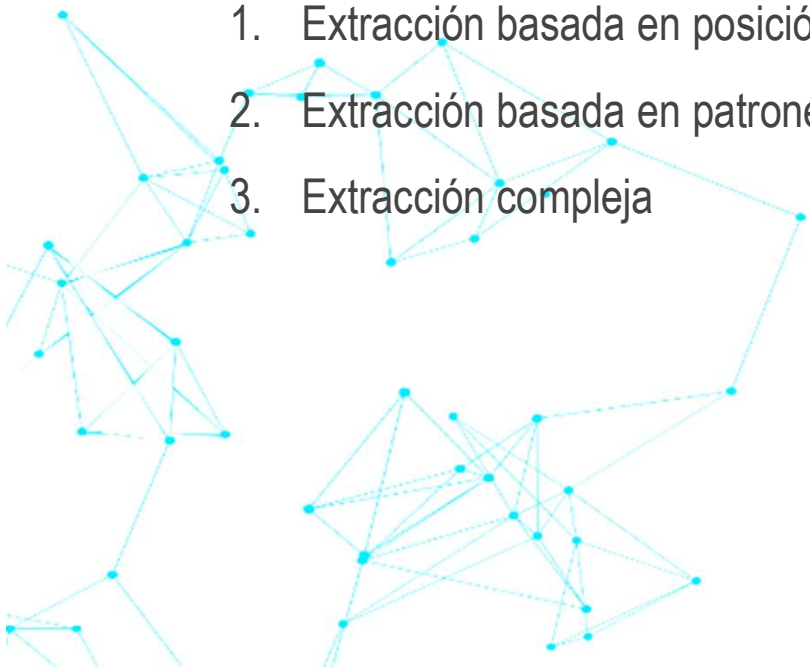
Cambio de granularidad de los datos mediante agregaciones o pivote de campos.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*intra-registro*)

Extracción de valores: Consiste en crear un nuevo campo a partir de uno ya existente. Por ejemplo, identificar y extraer subcadenas de los registros de una columna para crear una nueva.

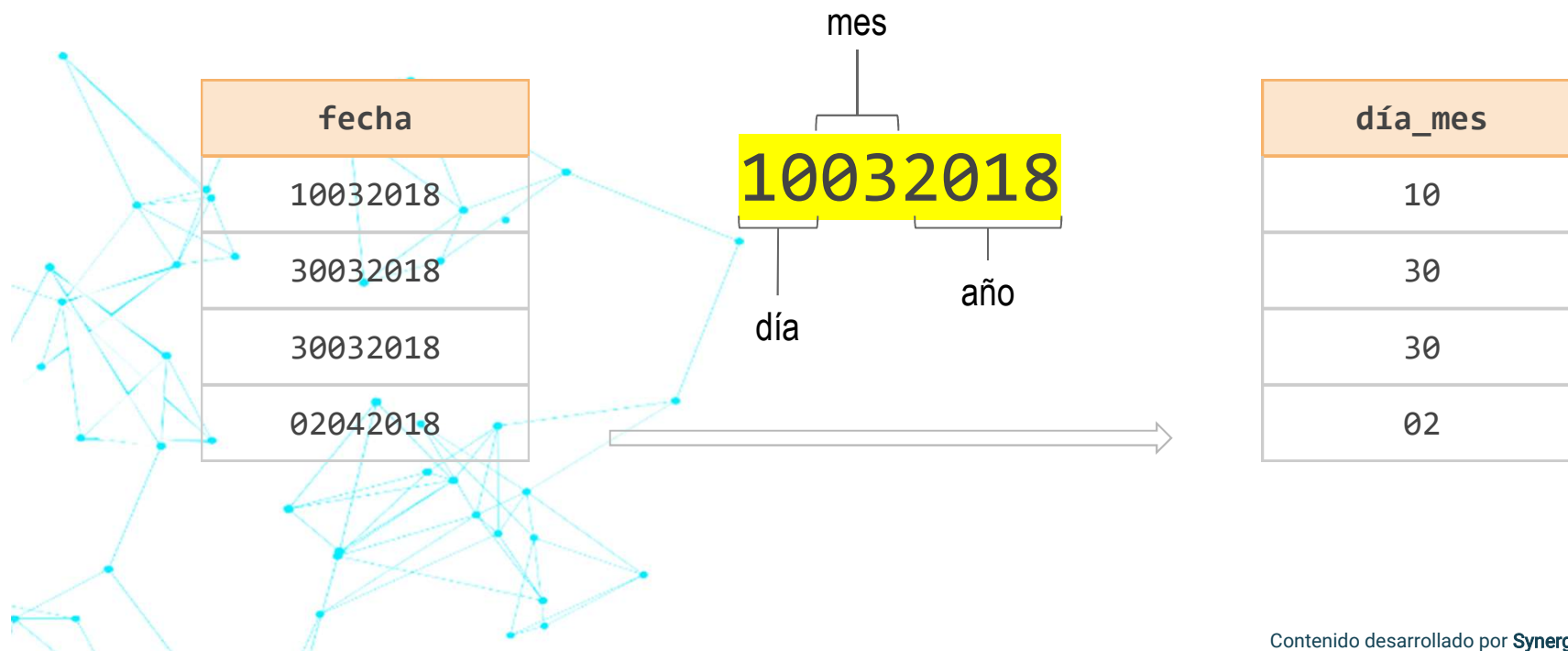
- 
1. Extracción basada en posición
 2. Extracción basada en patrones
 3. Extracción compleja

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*intra-registro*)

1. *Extracción basada en posición.* La extracción se realiza especificando la posición inicial y final de la subcadena de caracteres.

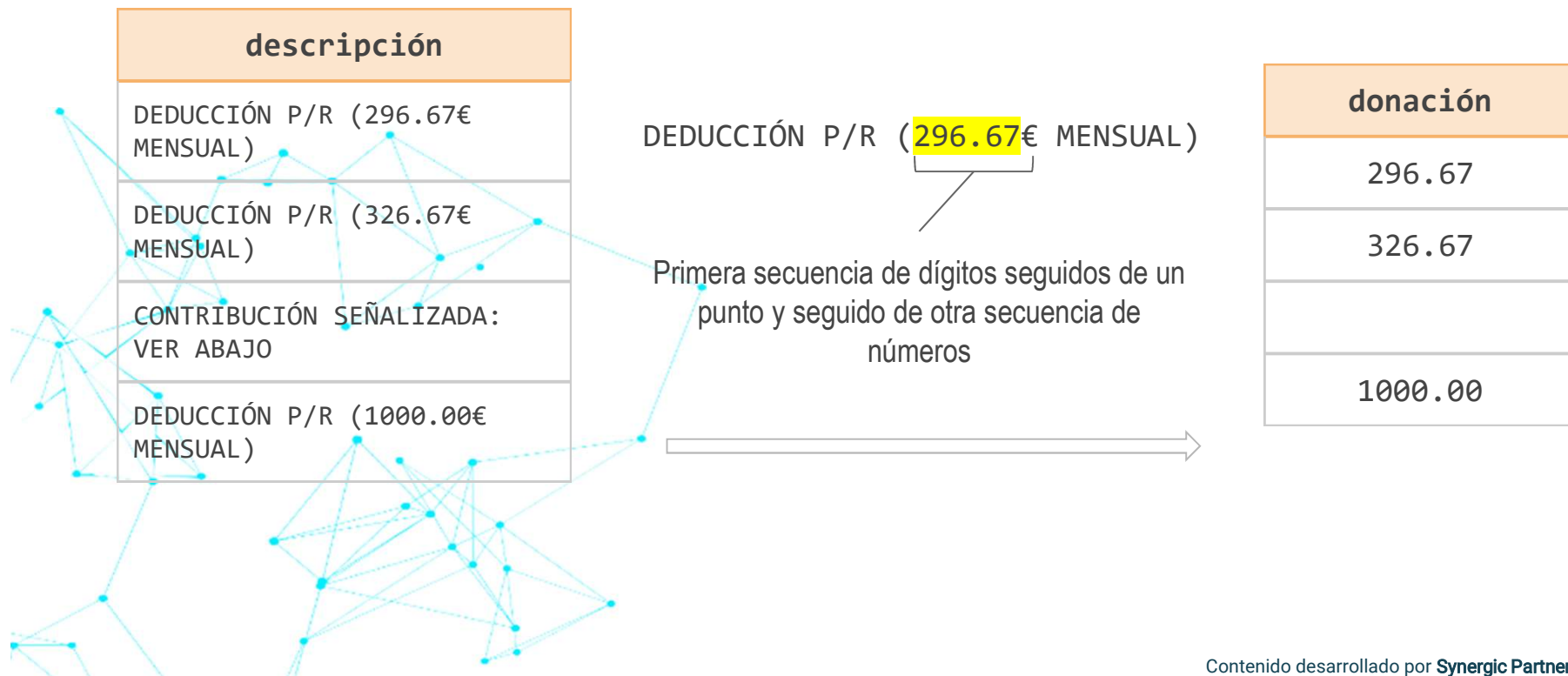


CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*intra-registro*)

2. *Extracción basada en patrones.* El método de extracción consiste en un conjunto de reglas que describen la secuencia de caracteres que se desea extraer.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*intra-registro*)

3. *Extracción compleja.* Consiste en extraer datos de estructuras jerárquicas complejas presentes en formatos semi-estructurados. Por ejemplo:

JSON array

Secuencia ordenada de valores

```
[“María”, “Pedro”, “Rosa”]
```

JSON map

Conjunto de pares clave-valor

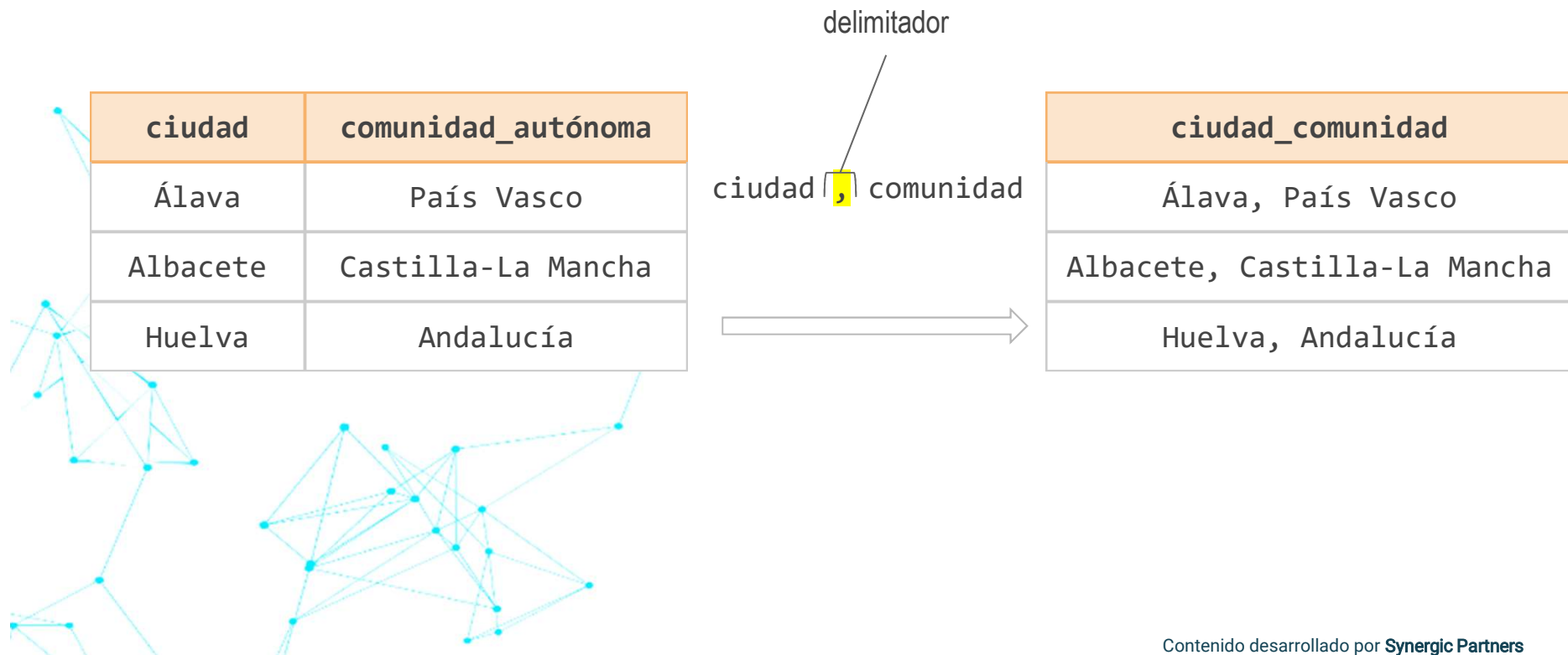
```
{  
  “producto”: “Trifacta Wrangler”,  
  “precio”: “gratis”,  
  “categoría”: “herramienta de data  
  wrangling”  
}
```

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*intra-registro*)

Combinación de campos: Consiste en fusionar los registros de uno o más campos relacionados.

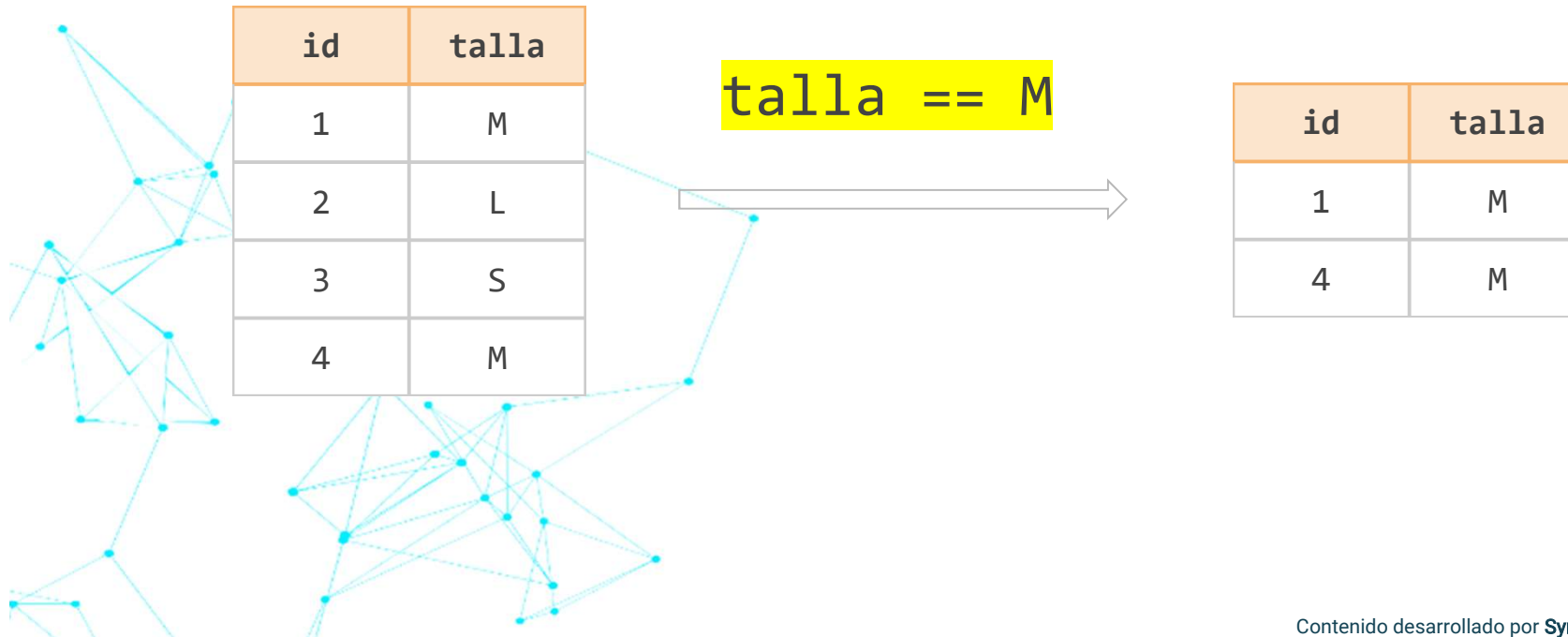


CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*inter-registro*)

Filtrado: Consiste en remover registros o campos enteros. En general se utiliza durante la limpieza, pero también se puede utilizar para cambiar la granularidad del dataset.



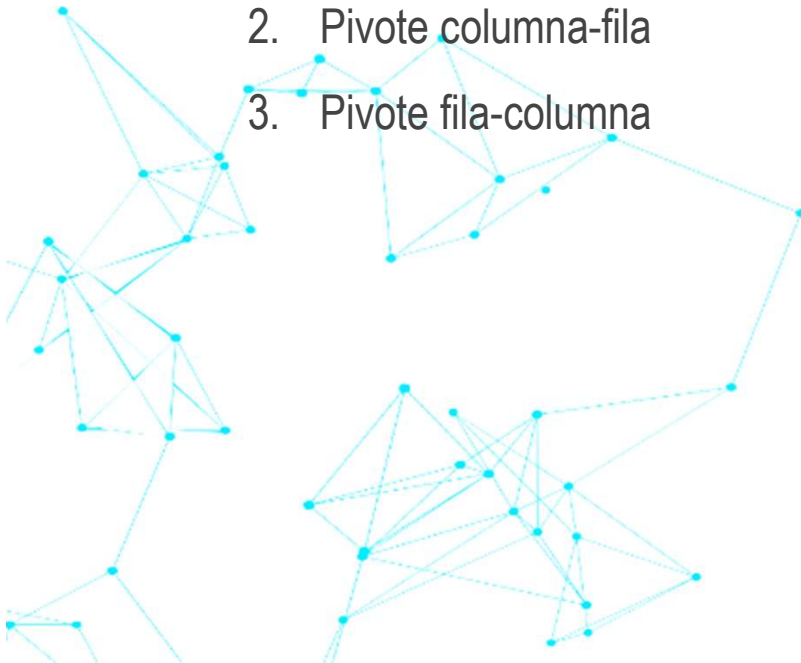
CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*inter-registro*)

Agregación / Pivote. Son operaciones que permiten cambiar la granularidad del dataset.

1. Agregaciones simples
2. Pivote columna-fila
3. Pivote fila-columna

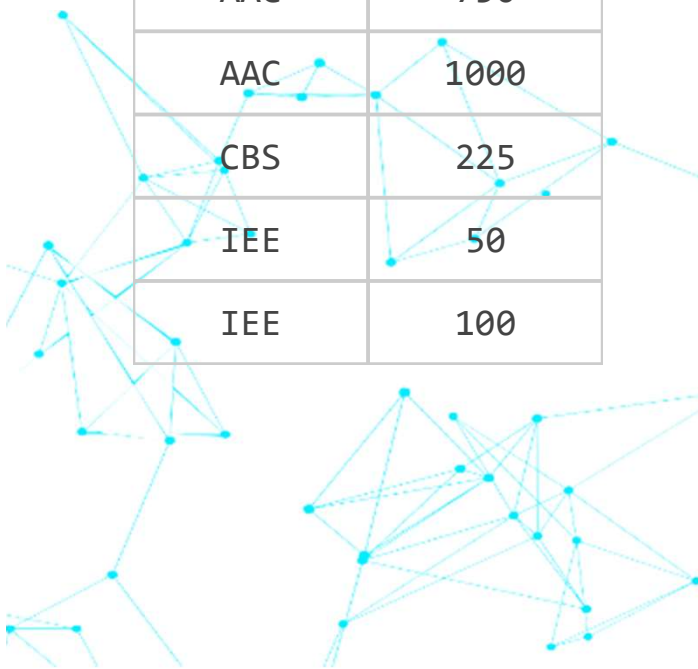


CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*inter-registro*)

1. *Agregaciones simples.* Consiste en mapear un conjunto de registros de entrada a uno de salida utilizando operaciones simples como: suma, media, mínimo, concatenación, etc.



empresa	donación
AAC	750
AAC	1000
CBS	225
IEE	50
IEE	100

por empresa

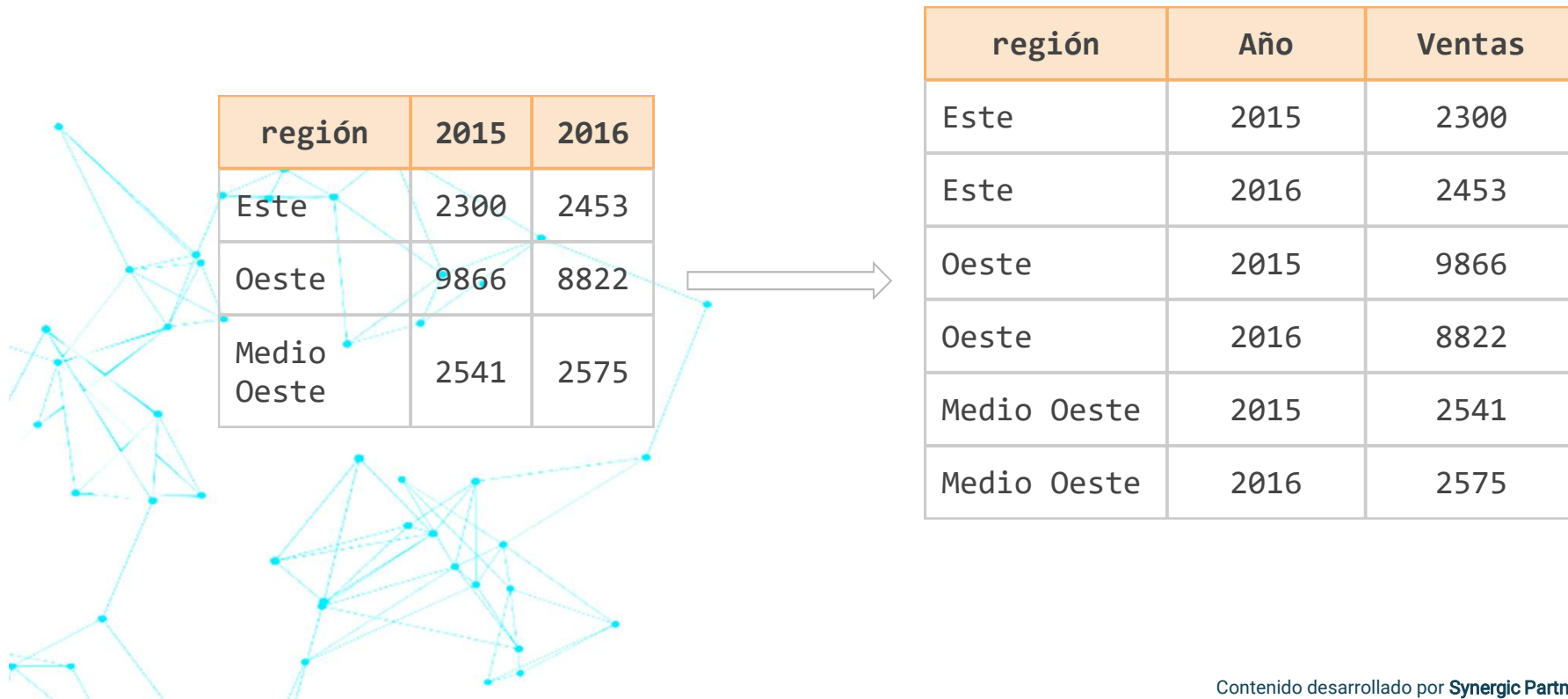
empresa	Suma de donaciones	Media de donaciones	Número de donaciones
AAC	1750	875	2
CBS	225	225	1
IEE	150	75	2

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*inter-registro*)

2. *Pivote columna-fila.* También se conoce como desnormalización. Es útil cuando los datos contienen columnas que representan el mismo tipo de datos, y se quieren resumir los datos más fácilmente.

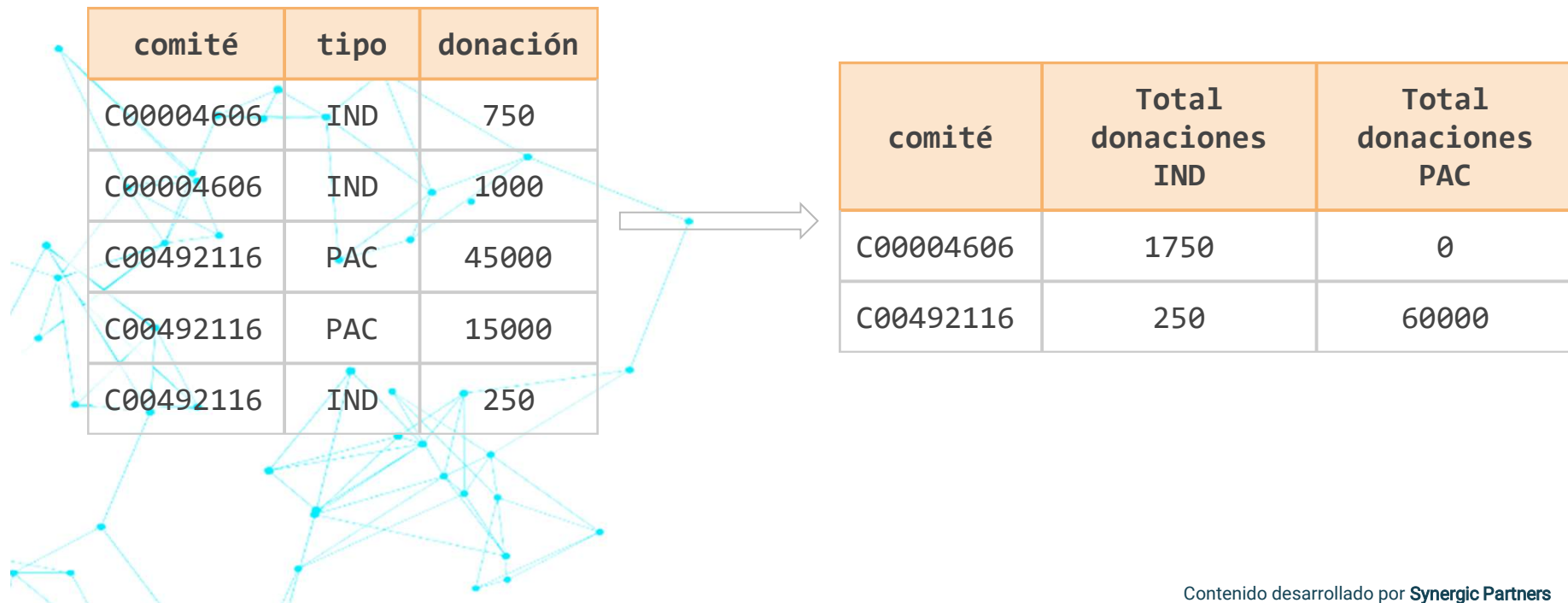


CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Estructurar (*inter-registro*)

3. *Pivote fila-columna*. Consiste en crear nuevos campos alrededor de una columna pivote, usando los registros únicos de una columna como etiquetas y realizando agregaciones simples o transformaciones más complejas en las columnas restantes.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Corresponde a las acciones que permiten incorporar nueva información al dataset. Puede consistir en insertar registros o campos adicionales desde otros datasets relacionados, o usar fórmulas para crear nuevos campos.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Unión. Consiste en adjuntar registros adicionales a un dataset ya existente; en otras palabras, consiste en crear un nuevo dataset al apilar verticalmente dos datasets relacionados.

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

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

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

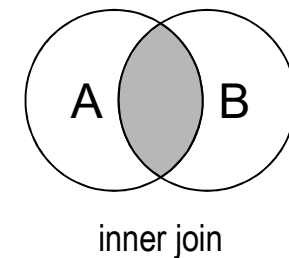
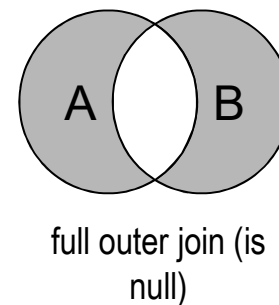
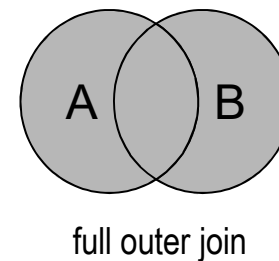
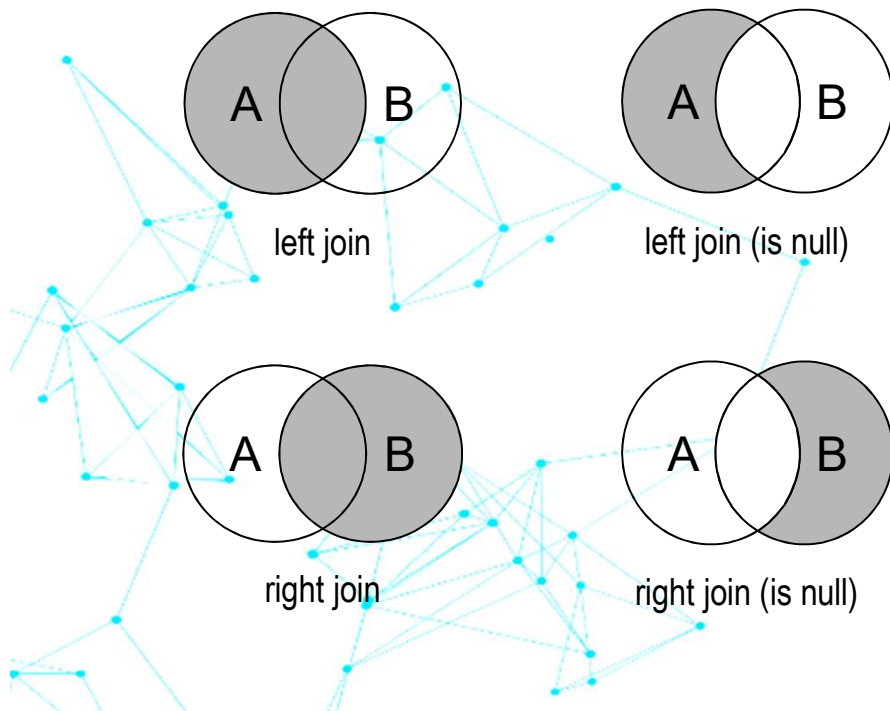
CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

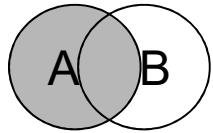
Joins. Los joins se usan para combinar filas de dos o más tablas, basados en un campo común (key field)

Supongamos que tenemos dos tablas, A y B, existen siete tipos de joins:

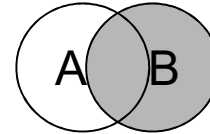


CONCEPTOS BÁSICOS DE DATA WRANGLING

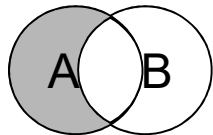
Acciones básicas. Transformación



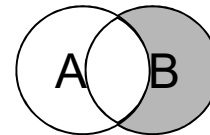
Left join. Devuelve todas las filas de A, junto con las de B que cumplen la condición.



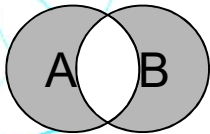
Right join. Devuelve todas las filas de B, y las de A que cumplen la condición.



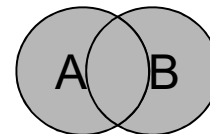
Left join (is null). Devuelve todas las filas de A, salvo las de B que no cumplen la condición.



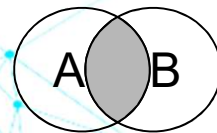
Right join (is null). Devuelve todas las filas de B, salvo las de A que no cumplen la condición.



Full outer join (is null). Devuelve todas las filas de A y B, salvo las que cumplen la condición.



Full outer join. Devuelve todas las filas de A y B.



Inner join. Devuelve todos los registros de la tabla A y de la tabla B donde se cumple la condición.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Ejemplo. Supongamos que tenemos estos dos conjuntos de datos:

df_mascotas

cod_distrito	distrito	especie_canina	especie_felina
2	Arganzuela	10591	3202
1	Centro	15470	6164
5	Chamartín	11759	2809
7	Chamberí	2979	13461
9	Moncloa Aravaca	12600	13461
3	Retiro	8183	2368
4	Salamanca	12709	3424
6	Tetuán	12427	3424
10	Latina	5476	1298

df_parques


barrio	cod_distrito	distrito	n_parques
Imperial	2	Arganzuela	1
Acacias	2	Arganzuela	1
Legazpi	2	Arganzuela	1
Delicias	2	Arganzuela	1
Chopera	2	Arganzuela	0
Palos de Moguer	2	Arganzuela	0
Atocha	2	Arganzuela	0
Palacio	1	Centro	1
Justicia	1	Centro	1
Vista Alegre	11	Carabanche 1	1

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

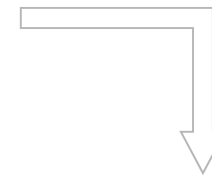
Enriquecer

Ejemplo (cont.). Queremos saber, para cada distrito, el total de parques que hay y el número de animales domésticos registrados en ellos. Primero, agrupamos los datos por **cod_distrito**, y modificamos **n_parques** para que sea la suma por barrios.



df_parques

barrio	cod_distrito	distrito	n_parques
Imperial	2	Arganzuela	1
Acacias	2	Arganzuela	1
Legazpi	2	Arganzuela	1
Delicias	2	Arganzuela	0
Chopera	2	Arganzuela	0
Palos de Moguer	2	Arganzuela	0
Atocha	2	Arganzuela	0
Palacio	1	Centro	1
Justicia	1	Centro	1
Embajadores	1	Centro	0



df_parques_agg

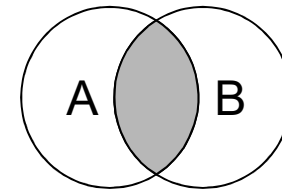
cod_distrito	distrito	n_parques
1	Centro	2
2	Arganzuela	4
11	Carabanchel	1

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Ejemplo (cont.). Si hacemos inner join:



inner join

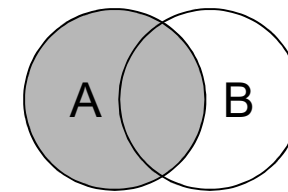
cod_distrito	distrito	n_parques	especie_canina	especie_felina
2	Arganzuela	4	10591	3202
1	Centro	2	15470	6164

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Ejemplo (cont.). Si hacemos left join:



left join

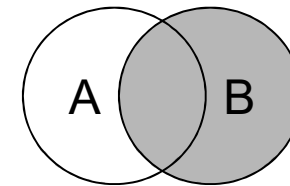
cod_distrito	distrito	n_parques	especie_canina	especie_felina
2	Arganzuela	4	10591	3202
1	Centro	2	15470	6164
5	Chamartín	NaN	11759	2809
7	Chamberí	NaN	2979	13461
9	Moncloa Aravaca	NaN	12600	13461
3	Retiro	NaN	8183	2368
4	Salamanca	NaN	12709	3424
6	Tetuán	NaN	12427	3424
10	Latina	Nan	5476	1298

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Ejemplo (cont.). Si hacemos right join:



right join

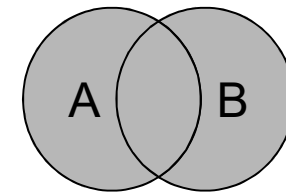
cod_distrito	distrito	n_parques	especie_canina	especie_felina
2	Arganzuela	4	10591	3202
1	Centro	2	15470	6164
11	Carabanchel	1	NaN	NaN

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Ejemplo (cont.). Si hacemos outer join:



full outer join


cod_distrito	distrito	n_parques	especie_canina	especie_felina
2	Arganzuela	4	10591	3202
1	Centro	2	15470	6164
5	Chamartín	NaN	11759	2809
7	Chamberí	NaN	2979	13461
9	Moncloa Aravaca	NaN	12600	13461
3	Retiro	NaN	8183	2368
4	Salamanca	NaN	12709	3424
6	Tetuán	NaN	12427	3424
10	Latina	Nan	5476	1298
11	Carabanchel	1	NaN	NaN

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Insertar metadatos. Consiste en añadir los metadatos al dataset. Los metadatos más comunes son: los nombres de los ficheros fuentes, números de registros, fecha y hora de creación/actualización/acceso, etc.



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

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Enriquecer

Variables derivadas. Consiste en generar nuevos valores. Existen dos tipos: genéricas y propietarias.



Día de la semana o estación a partir de la fecha.



Código postal, coordenadas de latitud/longitud a partir de una dirección.



Modelos customizados específicos de una organización.

CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

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.



CONCEPTOS BÁSICOS DE DATA WRANGLING

Acciones básicas. Transformación

Limpiar

- ✓ **Valores ausentes/nulos**
- ✓ **Valores inválidos.** Son valores inconsistentes con otros campos, ambiguos, o con mal codificados.



Filtrado de valores nulos.



Imputación de valores ausentes, usando la media, mediana, etc.



Marcar valores inválidos

Telefónica
FUNDACIÓN

Conecta Empleo

