



## Data Management

### KeepCoding

**Fechas: 12 y 13 de diciembre de 2019.**

**Horas: 8h**

**Instructora: Nerea Sevilla**

**Nombre Material: Ejercicios\_DataCleaning\_Talend\_Preparation**

## Data Cleansing con Talend Preparation

### Ejercicio 1.-Data Cleansing de Contactos de Clientes

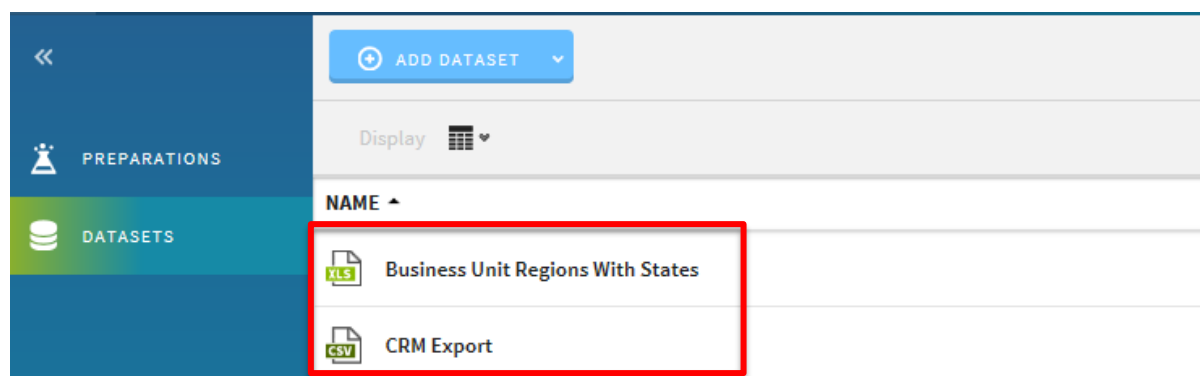
Talend Preparation es una interfaz web muy fácil de usar, que proporciona la capacidad de autoservicio de calidad de datos y enriquecimiento de datos antes del análisis, a los usuarios y analistas del negocio.

En los ejercicios aplicaremos algunas de las técnicas de la evaluación y resolución de problemas de calidad de datos visto en la parte teórica.

- Perfilado para descubrir frecuencias y formatos de datos.
- Data Cleansing de datos inválidos en las dimensiones de validez y consistencia en datos.
- Detección de valores atípicos.
- Estandarización de datos
- Enriquecimiento de datos, mediante la creación de nuevas columnas basadas en otras.
- Enmascaramiento u ofuscación de datos.
- Fusión de datos
- Agrupaciones de datos.

### Importación de Data Set

La pantalla de DATA SET de Talend, nos presenta los dataset que ya están importados, con los que estemos trabajando o la posibilidad de importar más ficheros.



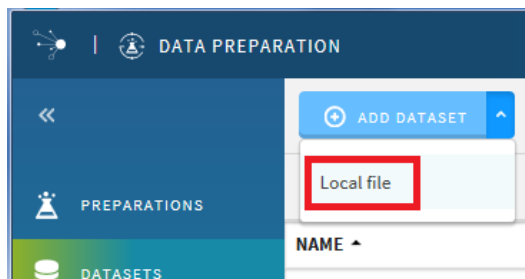
En la misma pantalla nos da la posibilidad de importar un nuevo fichero local. En la versión gratuita sólo se puede importar datos desde ficheros locales, en la versión de pago se puede importar desde servidores remotos como una base de datos, un Amazon Web Service o un servicio cloud como Salesforce)



Vamos a traer los datos de un fichero local Potenciales\_Clientes.xls que os hemos entregado como parte del material del curso y ver si los podemos mejorar su calidad.

Somos gerentes de una empresa de estudio de mercados y hemos recibido nuestro socio los datos obtenidos de un evento en un Excel, que incluye datos valiosos de información de contacto de clientes y su clasificación como potenciales clientes para campañas de marketing. Es necesario descubrir, limpiar y preparar la información sobre clientes potenciales antes de poder analizar el retorno de la inversión del evento e integrar los datos de prospección en otros sistemas de información de nuestra empresa, como CRM.

Para ello importaremos el fichero, utilizando la funcionalidad ADD DATASET, que se encuentra en la parte superior en la sección de DataSets.



### Profiling Automático con Talend.

En la teoría vimos que la primera operación a realizar para evaluar la calidad de los datos que debe realizarse es un perfilado o profiling, para determinar cuánto de buenos o malos son nuestros datos y obtener unas estadísticas sobre los datos. Talend Preparation, nos ayuda muchísimo en esta tarea, ya que tienen la capacidad de realizar profiling a nivel de columna, fila y tabla de los datos.

Talend una vez importados los datos, nos presenta una tabla con los datos de los clientes y los nombres de las columnas que detecta automáticamente al importar, con el tipo de campo en función de los datos que incluye cada columna (relaciona el tipo del campo con la semántica del campo) basándose en la correlación de datos.



DATA PREPARATION

Cientes

Filters 6040/6040

Add a filter...

	Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out	Salary_Out	Address	City
	integer	First Name	text	Gender	text	text	text	text	Address Line	
1	1	James	Butt	F	Under 18	K-12 Student	Single	0	6649 N Blue Gun St	Ni
2	2	Josephine	Darakjy	M	56+	Self-Employed	Married	100,000-149,999	4 B Blue Ridge Blvd	Bl
3	3	ART	Venere	M	25-34	Scientist	Married	< 50,000	8 W Cerritos Ave #54	Br
4	4	Lenna	Paprocki	M	45-49	Executive/Managerial	Divorced	150,000-199,999	639 Main St	Ar
5	5	Donette	Foller	M	25-34	Writer		50,000-99,999	34 Center St	Hi
6	6	Simona	Morasca	F	50-55	Homemaker	Married	100,000-149,999	3 McAuley Dr	Ar
7	7	Mitsue	Tollner	M	35-44	Academic/Educator	Divorced	100,000-149,999	7 Eads St	Cl
8	8	Leota	dilliard	M	25-34	Programmer		100,000-149,999	7 W Jackson Blvd	Sc
9	9	Sage	Wieser	M	25-34	Technical/Engineer	Divorced	150,000-199,999	5 Boston Ave #88	S
10	10	krisj	Marrier	F	35-44	Academic/Educator	Divorced	< 50,000	228 Runamuck Pl #288	Br
11	11	minna	Amigon	F	25-34	Academic/Educator	Divorced	150,000-199,999	2371 Jerrold Ave	Ki
12	12	Abel	Maclead	M	25-34	Programmer	Divorced	< 50,000	37275 St Rt 17m	M
13	13	Kiley	Calderera	M	45-49	Academic/Educator		150,000-199,999	25 E 75th St #69	Li
14	14	Graciela	Ruta	M	35-44	Other	Divorced	< 50,000	98 Connecticut Ave N	Cl
15	15	Commy	Albares	M	25-34	Executive/Managerial		> 200,000	56 E Morehead St	Li
16	16	Mattie	Poquette	F	35-44	Other		< 50,000	73 State Road 434 E	Pl
17	17	Meaghan	Garufi	M	50-55	Academic/Educator	Divorced	> 200,000	69734 E Carrillo St	M
18	18	Gladys	Rim	F	18-24	Clerical/Admin		50,000-99,999	322 New Horizon Blvd	M
19	19	Yuki	Whobrey	M	Under 18	K-12 Student	Single	0	1 State Route 27	Ti
20	20	Fletcher	Flosi	M	25-34	Sales/Marketing	Divorced	> 200,000	394 Manchester Blvd	Ri

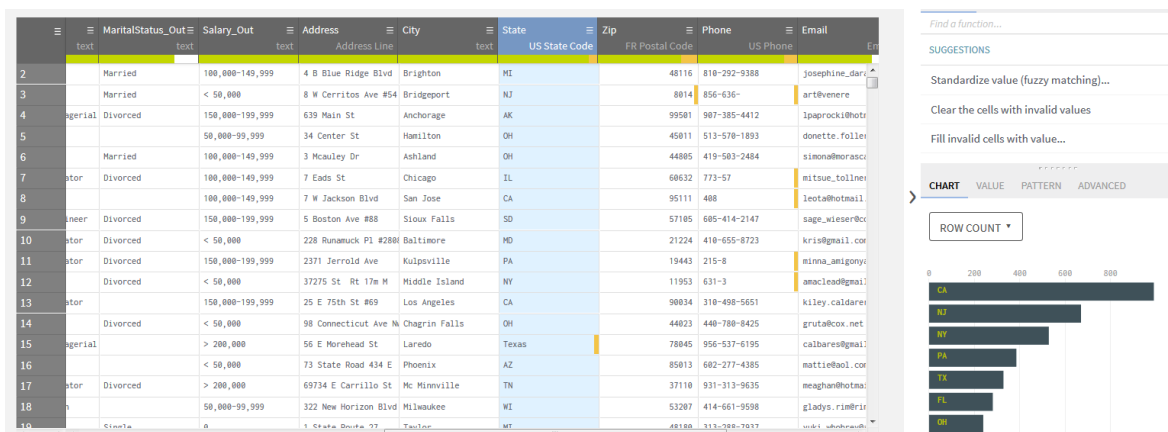
Por ejemplo, la columna STATE ha sido identificada como estados de EEUU,

ny	city	state	date	campaign_id	le
			date	text	
		This column is a US State Code			
		US State Code 94.33 %			

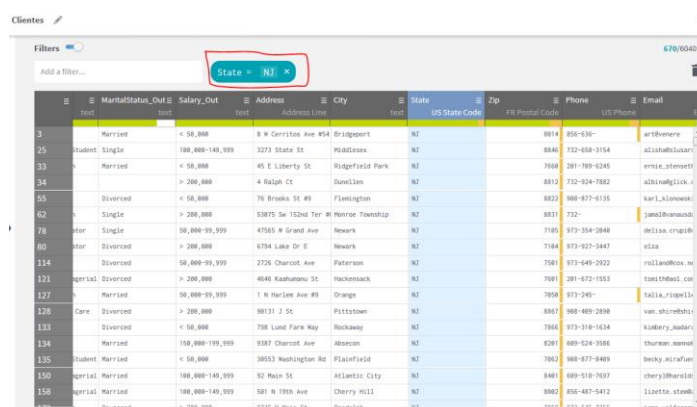
Al hacer clic en la columna de estado,

State
US State Code
MI
NJ
AK
OH

La preparación de datos proporciona una representación visual de mis datos. En la esquina inferior derecha ofrece los datos de distribución por estados de EEUU. Si estuviéramos en una versión de pago, incluso nos mostraría el mapa de EEUU con la distribución de los datos



Si hago clic en un estado dentro de una barra del gráfico, crea un filtro en mis datos que me permite ver solo los registros de ese estado.



Al igual que con Trifacta, las acciones que vayamos realizando para la preparación de los datos son pasos que se van añadiendo a la receta de preparación. Y nos podemos mover por los pasos, eliminando o activándolos según vayamos necesitando.

Al final la preparación de los datos, consistirá en la aplicación de una serie de pasos que son las acciones realizadas hasta conseguir el resultado final que cumplan con los mínimos de calidad para poder ser utilizados.

## Cleansing con Talend.

Vamos a comenzar con la limpieza de algunos datos.

El profiling de Talend, nos proporciona la barra de calidad, que aparece debajo del encabezado de cada columna. Muestra la cantidad de registros correctos, vacíos e incorrectos que tiene dicha columna. Cada uno de ellos se representa por un color.

1. Si es verde, la cantidad de datos correctos. Los datos coinciden con el formato de la celda.
2. Si es naranja, la cantidad de datos incorrectos. Los datos no coinciden con el formato de la celda.
3. Si es blanco, datos vacíos.

Solucionemos los registros inválidos del campo STATE, haciendo clic en la opción de Valores Inválidos.



city state

Airport US State Code

Select rows with invalid values for state

Potenciales\_Clientes

Filters

state : rows with invalid values

id	Name	last_name	email	job_title	company	city	state
integer	First Name	Last Name	Email	text	text	Airport	US State Code
213	Edward	Kennedy	ekennedy5@youtu.be	Executive Secretary	Flipopia	Cedar Rapid	pi
754	Anne	Matthews	amattewsg0@soup.io	Staff Scientist	Eazzy	Dallas	Texas
756	Justin	Lopez	jlopezio@geocities.jp	Clinical Specialist	Flipstorm	Austin	Texas
757	Eric	Crawford	ecrawfordj@nasa.gov	Administrative Assist	Ozu	Dallas	Texas
765	Jennifer	Shaw	jshawpm@uiuc.edu	Occupational Therapist	Roexo	Plano	Texas
961	Ralph	Webb	rwebbrk@theguardian.co.uk	Administrative Assistant	Thoughtmix	Dallas	Texas
985		Walker					E

En relación a la dimensión de consistencia de la calidad de los datos, vista en la parte de teoría. Existe un registro inconsistente. Ya que hay inconsistencia en la relación del campo Ciudad con el campo estados.

Cedar Rapids está en Iowa. Sustituimos el valor que indica por IA y aplicamos el cambio a todas las celdas.

city	state	date
Airport	US State Code	date
Cedar Rapid	IA	7/28/2015
Dallas		
Justin		

☒ Apply to all cells with this value

En relación a la dimensión de validez de la calidad de los datos, hay 5 registros inválidos porque no siguen el formato estándar para el estado que es la abreviatura de 2 caracteres.

Vamos a corregir estos registros

El estado Texas debe ser abreviado para coincidir con el formato del resto de registros, la modificación se aplica a las otras celdas con el mismo valor.

state	date
US State Code	date
TX	11/13/2015

☒ Apply to all cells with this value

Por último, nos queda un registro invalido, parece que se añadió el valor "E" por error teniendo en cuenta que la columna de ciudad no tiene valor.



city	state
Airport	US State Code
	E

Ahora ya está el campo estado (state) sin ningún registro inválido, el color naranja ha desaparecido y en la barra de calidad de datos sólo hay color verde y blanco.

Tampoco queremos emails inválidos dentro de los datos importados, por lo tanto, los eliminaremos.

last_name	email
Last Name	Email
Select rows with invalid values for email	
Clear the cells with invalid values	
Delete the rows with invalid cell	

Eliminar registros no válidos al igual que cuando se hace cualquier cambio en la preparación de datos, no altera los datos originales.

La columna Nombre (name) también presenta inconsistencias porque hay casos que aparecen los datos en mayúsculas y otras en minúsculas y además hay algunos espacios en blanco en el campo nombre. Como queremos eliminar espacios en blanco del nombre, hacemos clic en el campo nombre y seleccionamos la sugerencia “Remove trailing and leading characters” para eliminar los espacios en blanco y pulsamos el botón SUBMIT, sino aparece como sugerencia tecleamos en la casilla de búsqueda, los primeros caracteres de la función, hasta que aparezca.

Name	last_name	email	job_title	company	city	state	date
First Name	Last Name	Email	text	text	Airport	US State Code	
Kathryn	Garcia	kgarcia148g					
Jason	Alexander	jalexander44@gmail.c	Chemical Engineer	Abata	Pearl City	HI	22/1
Lillian	Simpson	lsimpsonf7@gmail.com	Desktop Support Tech	Camimbo	Wichita	KS	2/28
WALTER	Ruiz	wruiz1z@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK	7/15
Joshua	Hunt	jhuntek@last.fm	Financial Advisor	Oyope	Wilmington	DE	3/16
Mildred	Flores	mflores06@earthlink.	Nurse	Edgeblab	Miami	FL	10/1
Victor	Gonzalez	vgonzalez8@npr.org	Sales Associate	Ntag	Atlanta	GA	17-1
Joshua	Simmons	jsimmons5@newyorker	Occupational Therapi	Oba	Jacksonville	FL	17-1
Beverly	Wright	bwright3@arizona.ed	Biostatistician	Skynoodle	Indianapolis	IN	01/0
Fred	Rodriguez	frrodrigueznc@fotki.c	Director of Sales	Eidel	Anchorage	AK	7/6/
Joseph	Peterson	jpetersonn@sohu.com	Research Nurse	Gabcube	Las Vegas	NV	3/16

remove

Remove trailing and leading characters...

☐ Create new column

Padding character:


Whitespace

SUBMIT



STRINGS ADVANCED

En la parte izquierda de la pantalla aparece la RECETA de la preparación, es decir los diferentes pasos que estamos aplicando sobre los datos para limpiar los datos.





Potenciales\_Clientes Preparation 



---

1 Search and replace on column state  


---

2 Search and replace on column state  


---

3 Replace value on cell  


---


4 Delete the rows with invalid cell on column email 

---

5 Remove trailing and leading characters on column Name 

☐ Create new column

Padding character:  
Whitespace 



SUBMIT

Si se decide que ya no se quiere aplicar una determinada acción a los datos, se puede deshacer, haciendo clic en la papelera, también se pueden reorganizar las acciones y automáticamente se reflejan en los datos.

### Estandarización de datos.

Estandarizaremos de las columnas de Nombre (Name) y Apellido (last\_name) para que todos los registros sean en mayúsculas y eliminar las inconsistencias porque hay casos que aparecen los datos en mayúsculas y otras en minúsculas

Para ello realizaremos una selección múltiple de columnas. Haremos clic en el nombre y pulsando la tecla SHIFT, haremos clic sobre el apellido.

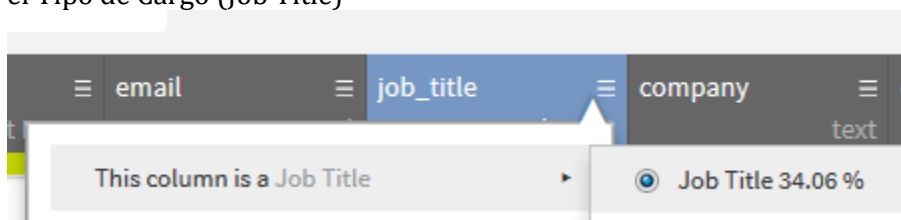
El siguiente paso es aplicar la función para cambiar a mayúsculas. Vamos a utilizar las funciones de la lista de sugerencias que aparecen a la derecha de la pantalla. Escribiendo el principio de la función que queremos utilizar, en nuestro caso, "Chang to upper case", la lista de sugerencias se adapta a lo que estamos escribiendo y se aplicaría a la columna o registro seleccionado, en nuestro caso a las columnas nombre y apellido, pulsaremos SUBMIT



Name	last_name	email	job_title	company	city	state	state_code
First Name	Last Name	Email	Text	Text	Text	Text	US State Code
JASON	Alexander	jalexander44@gmail.com	Chemical Engineer	Abata	Pearl City	HI	
LILLIAN	Simpson	lsimpson77@gmail.com	Desktop Support Tech	Camisbo	Wichita	KS	
WALTER	Ruiz	wruiz12@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK	
JOSHUA	Hunt	jhuntsk@last.fm	Financial Advisor	Oyope	Wilmington	DE	
MILDRED	Flores	mflores06@earthlink.net	Nurse	Edgebiab	Miami	FL	
VICTOR	Gonzalez	vgonzalez8c@npr.org	Sales Associate	Ntag	Atlanta	GA	
JOSHUA	Simmons	jsimmons59@newyorker.com	Occupational Therapist	Oba	Jacksonville	FL	
BEVERLY	Wright	bwright38@arizona.edu	Biostatistician	Sknoodle	Indianapolis	IN	
FRED	Rodriguez	frodrigueznc@fotki.com	Director of Sales	Eidel	Anchorage	AK	

El resultado será que todos los registros se actualizarán a datos con mayúsculas.

Talend asocia el tipo de dato en función del contenido para ayudarnos a descubrir tipos de datos, como hemos visto con el campo estado (STATE). Pero podemos cambiar esas sugerencias basándonos en nuestra propia experiencia. En el caso del campo Cargo (job\_title) Talend ha sugerido que el tipo de campo es Text. Pero podemos cambiarlo por otro más significativo, como el Tipo de Cargo (Job Title)



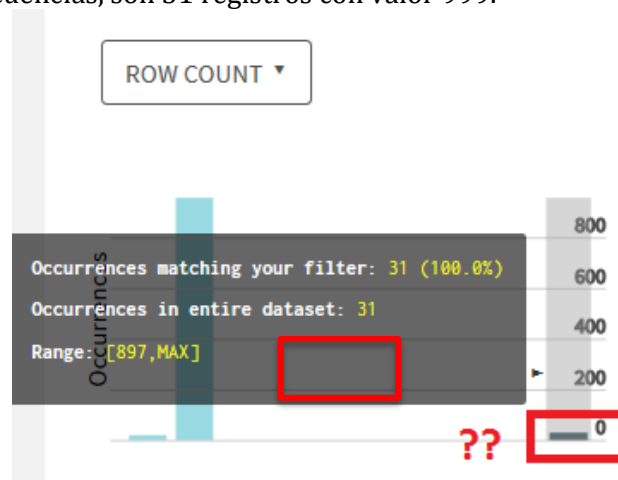
En versiones de pago, Talend permite crear tipos de campos semánticos personalizados más específicos para cada negocio, en mi caso tipos relacionados con el empleo, por ejemplo código de identificación fiscal según la primera letra del CIF (A, Sociedades Anónimas, Q, organismos autónomos, etc....)

El importante asociar los campos a tipos semánticos, porque Talend nos va a permitir realizar agrupaciones sobre estos campos, pudiendo enriquecer los datos para clasificar los datos evitando muchos tipos de datos diferentes.

### Detección de Valores Numéricos Atípicos..

Si seleccionamos el campo potencial (lead\_score) que se refiere a la puntuación asignada como potencial cliente en el futuro.

Al ser un campo numérico, el profiling nos aparece un histograma con la frecuencia de valores y vemos que aparece muy sesgado por algunos valores. Si seleccionamos sobre el histograma el grupo máximo de frecuencias, son 31 registros con valor 999.







Parace que se ha utilizado 999 como valor por defecto en ese campo, vamos a cambiarlo utilizando la función Fill cell with value, para ello tecleamos Fill y en la lista de sugerencias nos aparecerá automáticamente la función que necesitamos. Cambiaremos el valor 999 por 0.

COLUMN ROW TABLE

fill

Fill cells with value...

Use with:

Value

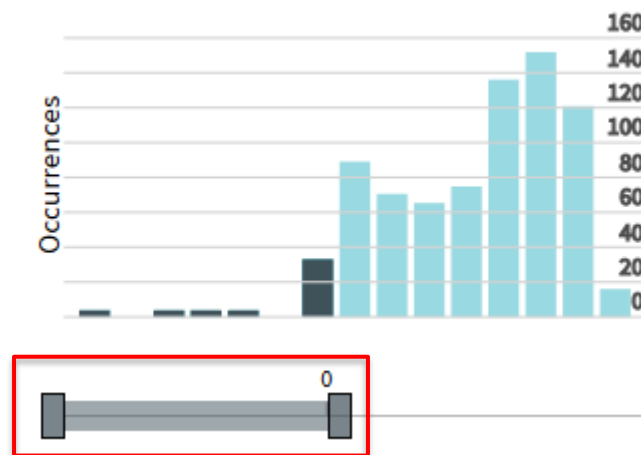
Value:

0

SUBMIT

De esta manera detectamos y limpiamos valores atípicos en campos numéricos, utilizando los gráficos para filtrar datos y cambiarlos.

Siguiendo con la misma columna, vemos que contiene valores negativos y no tiene sentido. Vamos a eliminarlos, primero en el histograma seleccionamos los valores negativos, moviendo la barra de valores del histograma.



Aplicaremos a los valores filtrados la función Calculate absolute value, bien tecleando la función o eligiendo la función de las sugerencias de Talend, si es que lo hace.

lead\_score

COLUMN ROW TABLE

calcu

Calculate absolute value...

☐ Create new column

SUBMIT

## Cleaning de Fechas.

Vamos a revisar los formatos de los datos de la columna date, para ello seleccionamos la columna fecha (date) y observamos sus formatos en la vista Pattern para poder observar mejor



los diferentes formatos de fechas y máscaras utilizados. Algunas fechas se han introducido con formato europeo (d/M/yyyy) y otras usan el formato americano (M/d/yyyy), y algunas usan el separador – y otras /. Para estandarizar las fechas utilizaremos el mismo formato para todas las fechas (dd/MM/yyyy) y la función **“Change date format”**, utilizando **Other** para especificar diseños personalizados.

### 13 Change date format on column date

☐ Create new column

Current format:  
I don't know, best guess

New format:  
Other

Your format:  
dd/MM/yyyy

Podemos cambiar el formato de fecha desde la Receta y para que en la fecha aparezca el nombre.

### 13 Change date format on column date

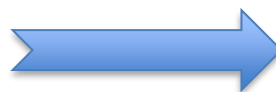
☐ Create new column

Current format:  
I don't know, best guess

New format:  
Other

Your format:  
dd.MMMM.yyyy

SUBMIT



date
22.November.2015
28.February.2015
15.July.2015
16.March.2015
15.October.2015
17.December.2014
17.December.2015
01.January.2016
06.July.2015
16.March.2015
09.December.2014
01.June.2015
02.November.2015
04.June.2015
02.November.2015
25.December.2014
31.August.2015



Ahora nos gustaría asociar cada registro a un nombre de campaña, para ello extraeremos datos para la creación de un nuevo campo desde el campo fecha (date), el año, utilizando la función “Extract date part”, seleccionando SOLO la opción de Year y SUBMIT para extraer.

The screenshot shows a data table with columns: date, campaign\_id, and lead\_score. The 'date' column contains full dates like '22.November.2015'. To the right, the 'date' configuration panel is open, showing the 'Extract date parts...' options. The 'Year' checkbox is checked, while 'Quarter of the year' and 'Month of the year' are unchecked.

date	campaign_id	lead_score
22.November.2015	HOCKEY_Y15Q01_cant	5
28.February.2015	RUN_Y14Q02_deal	36
15.July.2015	TRAIL_Y14Q04_purr	92
16.March.2015	HOCKEY_Y14Q02_mode	79
15.October.2015	HOCKEY_Y15Q04_chum	46
17.December.2014	TRAIL_Y15Q03_hold	85
17.December.2015	TRAIL_Y14Q03_moon	40
01.January.2016	TRAIL_Y15Q04_rosy	57

Ahora si seleccionamos el campo año que hemos extraído (date\_year) y realizamos una concatenación con la función concatenación (Concatenate with) con el Prefijo (que se añadirá al principio de cada registro) Campaña\_, en un campo nuevo

The screenshot shows the same data table, but now the 'date' column has been replaced by 'date\_YEAR', which contains only the year (e.g., '2015'). The 'Concatenate with...' configuration panel is open, showing the 'Concatenate with...' options. The 'Create new column' checkbox is checked. The 'Prefix' is set to 'Campaña\_'. The 'Use with' dropdown is set to 'Value'.

date_YEAR	campaign_id	lead_score
2015	HOCKEY_Y15Q01_cant	5
2015	RUN_Y14Q02_deal	36
2015	TRAIL_Y14Q04_purr	92
2015	HOCKEY_Y14Q02_mode	79
2015	HOCKEY_Y15Q04_chum	46
2014	TRAIL_Y15Q03_hold	85
2015	TRAIL_Y14Q03_moon	40
2016	TRAIL_Y15Q04_rosy	57

Obtenemos el nuevo campo con el nombre de la campaña, que podemos renombrar utilizando como etiqueta Nombre\_Campaña, seleccionando la columna y botón derecho para elegir “Rename Column”.

Ahora podemos borrar la columna date\_month que ya no necesitamos, con la misma operación. Seleccionar la columna y botón derecho para elegir la opción “Delete Column”.

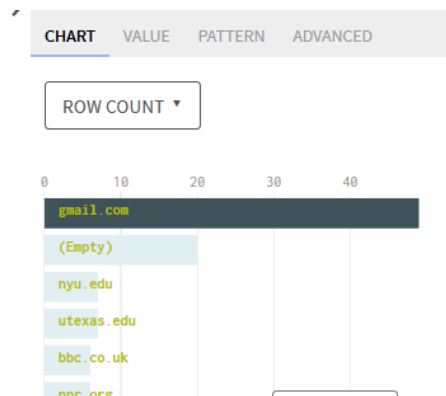
Haremos lo mismo para las columnas campaign\_id y id.

También sabemos que el departamento de ventas prefiere no usar dominios generales de emails como Gmail, así que vamos a eliminarlos. Para ello, extraeremos las diferentes partes que



componen el email y haremos un filtro por el dominio para eliminarlos. Utilizaremos la función **“Extract email parts”**.

Ahora en la gráfica seleccionamos, los registros del dominio Gmail



Y sobre el filtro, eliminamos los registros del dominio Gmail.

email_domain	job_title	company	city
Web Domain	Job Title		text
gmail.com	Chemical Engineer	Abata	
gmail.com	Desktop Support Tech	Camimbo	
gmail.com	Geological Engineer	Yakitri	
gmail.com	Professor	Blogpad	
gmail.com	Graphic Designer	Wikizz	

## Enmascaramiento u Ofuscación de Datos (Data Masking)

Cuando se manipula con datos de alto nivel según la clasificación de carácter privado de los datos, tales como nombres, direcciones, tarjetas de crédito o números de la seguridad social, la función de seguridad de la gestión de datos (la veremos más adelante) que define las normas que deben cumplir los datos para el correcto cumplimiento del Reglamento General de Protección de Datos, nos exige proteger los datos originales y deberemos aplicar técnicas de enmascaramiento u ofuscación de datos.

Vamos a aplicar data masking a los nombres de los clientes y aplicaremos la función Mask data (Ofuscación)

Los efectos del Data Masking son diferentes dependiendo del tipo semántico del campo donde se esté aplicando la función.

En el caso del nombre y apellido de los clientes que se han identificado como tipos de Talend First\_Name y Last\_Name, la función de ofuscación aplica una técnica y el resultado es diferente si se aplicará por ejemplo al campo email



First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out	Salary_Out	Action
Yzhqx	Yzhe	F	Under 18	K-12 Student	Single	0	
Dyyfypqjs	Akwppb	M	56+	Self-Employed	Married	100,000-149,999	
NAWRA	Benteu	M	25-34	Scientist	Married	< 50,000	
Resk1	Padzksrf	M	45-49	Executive/Managerial	Divorced	150,000-199,999	
Yaowzfh	Gmdjbu	M	25-34	Writer	Missing	50,000-99,999	

Mask

DATA MASKING

Mask data (obfuscation)

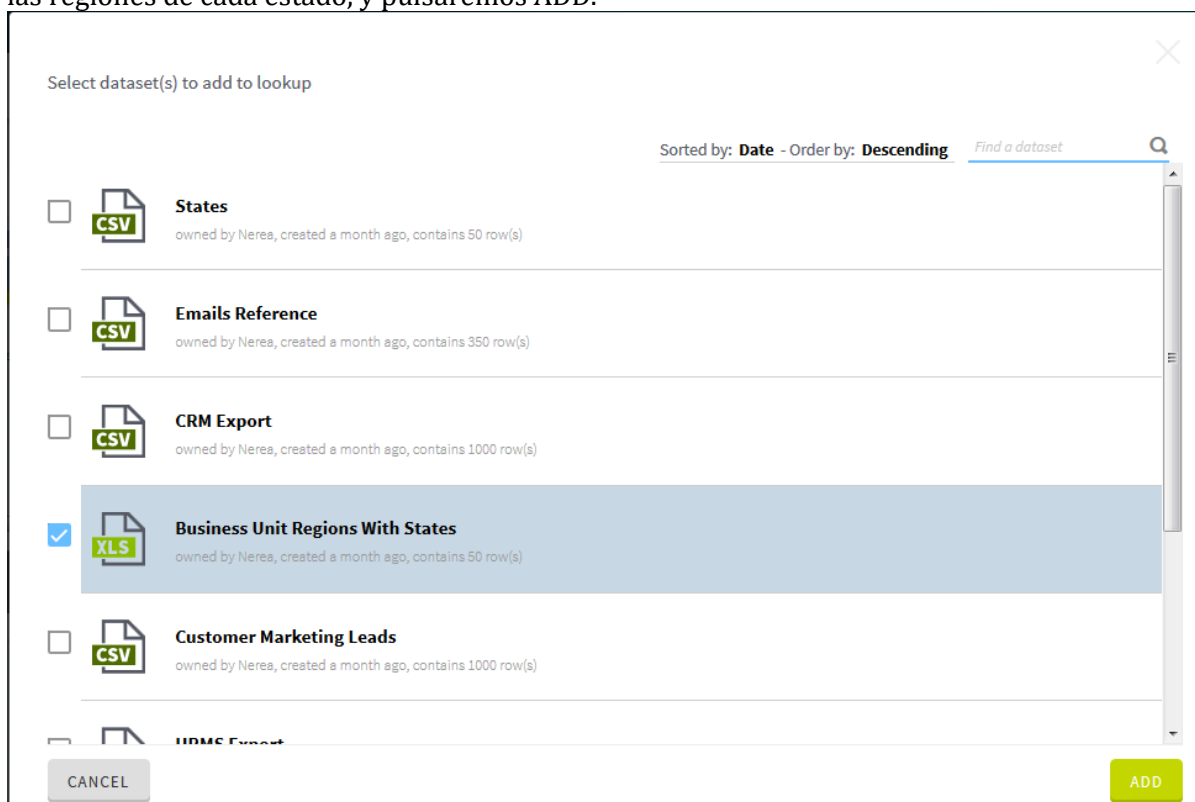
## Fusión de datos.

La fusión de datos en Talend se refiere a conectar datos desde diferentes fuentes. Esto permite coger datos desde otros dataset importados y añadirlos al dataset con el que se está trabajando.

Por ejemplo, vamos a utilizar dos dataset que vienen en la instalación de Talend Preparation. Customer y Business Unit Regions With States.

Abrimos el dataset Customer y pulsamos el botón Lookup

Y en la pantalla que nos aparece en la pantalla inferior pulsaremos + en la esquina inferior de la pantalla, y elegimos la fuente que queremos fusionar, en nuestro caso la fuente con las regiones de cada estado, y pulsaremos ADD.



Haremos clic con el campo de la fuente con la que estamos trabajando que queremos que utilice para unir los datos de la fuente destino y seleccionaremos el selector Add to DataSet y pulsaremos el botón CONFIRM.



Customers

Filters 6040/6040

Add a filter...

	Occupation	MaritalStatus_Out	Salary_Out	Address	City	State	Zip	Phone
	text	text	text	Address Line	text	US State Code	FR Postal Code	US Ph
1	K-12 Student	Single	0	6649 N Blue Gum St	New Orleans	LA	70116	504-621-8927
2	Self-Employed	Married	100,000-149,999	4 B Blue Ridge Blvd	Brighton	MI	48116	810-292-9388
3	Scientist	Married	< 50,000	8 W Cerritos Ave #54	Bridgeport	NJ	8014	856-636-
4	Executive/Managerial	Divorced	150,000-199,999	639 Main St	Anchorage	AK	99501	907-385-4412
5	Writer		50,000-99,999	34 Center St	Hamilton	OH	45011	513-570-1893
6	Homemaker	Married	100,000-149,999	3 McAuley Dr	Ashland	OH	44805	419-503-2484

State US State Code Region text

2	MT		Add to Dataset
3	OR	West	
4	ID	West	
5	WY	West	
6	CA	West	
7	NV	West	
8	UT	West	
9	CO	West	

ADD DATA FROM LOOKUP

1. Select two identical columns from two different datasets to link them. These columns turn blue.
2. Check "Add to Dataset" to select the columns you want to associate with the linked columns.
3. Place your mouse over the "Confirm" button to preview the result.

[Learn more...](#)

**CONFIRM**

**+** Business Unit Regions With...

Así hemos añadido información de las regiones a las que pertenece cada estado. Es importante hacer esta operación con los datos limpios para que el matching se haga correctamente.

State	Region
US State Code	text
MI	Mid West
NJ	North East
AK	West
OH	Mid West
OH	Mid West
IL	Mid West
CA	West
SD	Mid West
MD	North East
PA	North East
NY	North East
CA	West
OH	Mid West



## Agrupaciones de datos.

Las agrupaciones permiten descubrir registros en campos que tienen similar contenido y agruparlos juntos cambiando el texto para que coincida, para realizar segmentaciones en campañas de marketing.

Vamos a realizarlo sobre el campo Cargo (job\_title) y la función find and group similar text.

Al seleccionar el campo job\_title el gráfico de la parte inferior derecha nos muestra una gran cantidad de diferentes cargos. Para reducir la frecuencia de diferentes tipos de cargos, vamos a agrupar similares cargos en uno solo, mediante la agrupación.

The screenshot shows the KeepCoding interface. On the left, a table displays data for 'job\_title', 'company', 'city', and 'state'. The table has 1000/1000 rows. On the right, the 'job\_title' field is selected, and the sidebar shows the 'Find and group similar text...' option highlighted in red. Below the sidebar, a chart shows the frequency of different job titles.

job_title	company	city	state
Chemical Engineer	Abata	Pearl City	HI
Desktop Support Techn	Camimbo	Wichita	KS
Geological Engineer	Yakitri	Fairbanks	AK
Financial Advisor	Oyope	Wilmington	DE
Nurse	Edgeblab	Miami	FL
Sales Associate	Ntag	Altanta	GA
Occupational Therapist	Oba	Jacksonville	FL
Biostatistician	Skynoodle	Indianapolis	IN
Director of Sales	Eidel	Anchorage	AK
Research Nurse	Gabcube	Las Vegas	NV
Speech Pathologist	Zoomcast	Nampa	ID
Automation Specialist	Bluezoom	Bridgeport	CT
Automation Specialist	Shuffletag	Racine	WI
Librarian	Skalith	Bend	OR
Actuary	Rhyloo	Manhattan	NY
Senior Editor	Tazzy	Columbus	GA
Structural Engineer	Dynava	Overland Park	KS
Help Desk Operator	Gabtune	Orange	CT

job\_title

COLUMN ROW TABLE

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

Occupational Therapist

Database Administrator

VP Marketing

Financial Advisor

Web Designer

Software Consultant

Human Resources Manager

Librarian

Senior Financial Analyst

Geological Engineer

### FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

These values have been found	This value will be kept
<input checked="" type="checkbox"/> Administrative Assistant	Replace value: Administrative Assistant
<input checked="" type="checkbox"/> Administrative Officer	
<input checked="" type="checkbox"/> Senior Editor	Replace value: Senior Developer
<input checked="" type="checkbox"/> Senior Developer	
<input checked="" type="checkbox"/> Marketing Manager	Replace value: Marketing Manager
<input checked="" type="checkbox"/> Marketing Assistant	

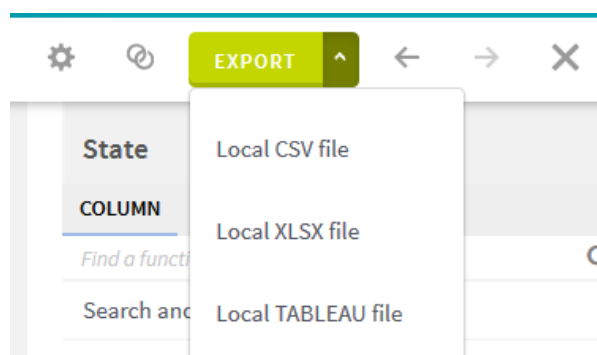
SUBMIT



En la primera columna de la pantalla con los grupos que ha generado Talend, son los elementos que van a ser agrupados, y en la segunda los grupos de clasificación de cargos propuestos. Se pueden deseleccionar los elementos o grupos que no interesen y también cambiar el nombre de los grupos. SUBMIT para agrupar.

### Exportación de Preparación.

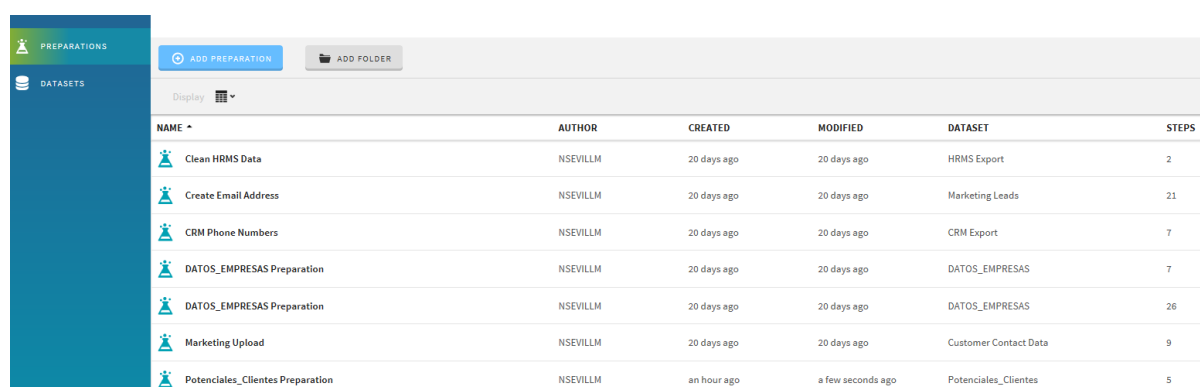
Cuando tengamos los datos completamente preparados, tenemos la opción de exportarlos. Se puede exportar a formato plano csv a formato Excel o formato Tableau



En versiones de pago, se ofrece más opciones de extracción en la nube, en bases de datos relacionales o Hadoop.

Como nuestro sistema CRM acepta ficheros XLS, lo exportamos en este formato.

La última tarea es guardar la preparación para poder aplicarla sobre otros datasets, le llamaremos Potenciales\_Clientes Preparation, al guardar se mostrará en la lista de preparaciones.



NAME	AUTHOR	CREATED	MODIFIED	DATASET	STEPS
Clean HRMS Data	NSEVILLM	20 days ago	20 days ago	HRMS Export	2
Create Email Address	NSEVILLM	20 days ago	20 days ago	Marketing Leads	21
CRM Phone Numbers	NSEVILLM	20 days ago	20 days ago	CRM Export	7
DATOS_EMPRESAS Preparation	NSEVILLM	20 days ago	20 days ago	DATOS_EMPRESAS	7
DATOS_EMPRESAS Preparation	NSEVILLM	20 days ago	20 days ago	DATOS_EMPRESAS	26
Marketing Upload	NSEVILLM	20 days ago	20 days ago	Customer Contact Data	9
Potenciales_Clientes Preparation	NSEVILLM	an hour ago	a few seconds ago	Potenciales_Clientes	5

El último paso es aplicar la preparación Potenciales\_Clientes Preparation a otros datasets. Imaginamos que pasado un tiempo se realiza otro evento para la captación de más clientes, lo que haremos es aplicar la preparación con los pasos que la componen al nuevo Data Set que nos han enviado, sin necesidad de volver a realizarla, utilizando el botón de preparation.



Ahora que los datos ya están limpios y estandarizados están **preparados** para:

**Análisis.** Por ejemplo, se pueden analizar las puntuaciones asignadas a los clientes por fecha y estado con Excel o un sistema de inteligencia de negocio.

**Integración.** Con los datos limpios, completos, consistentes y válidos están listos para integrar en un CRM o una aplicación automática de Marketing como Marketo o Salesforce.

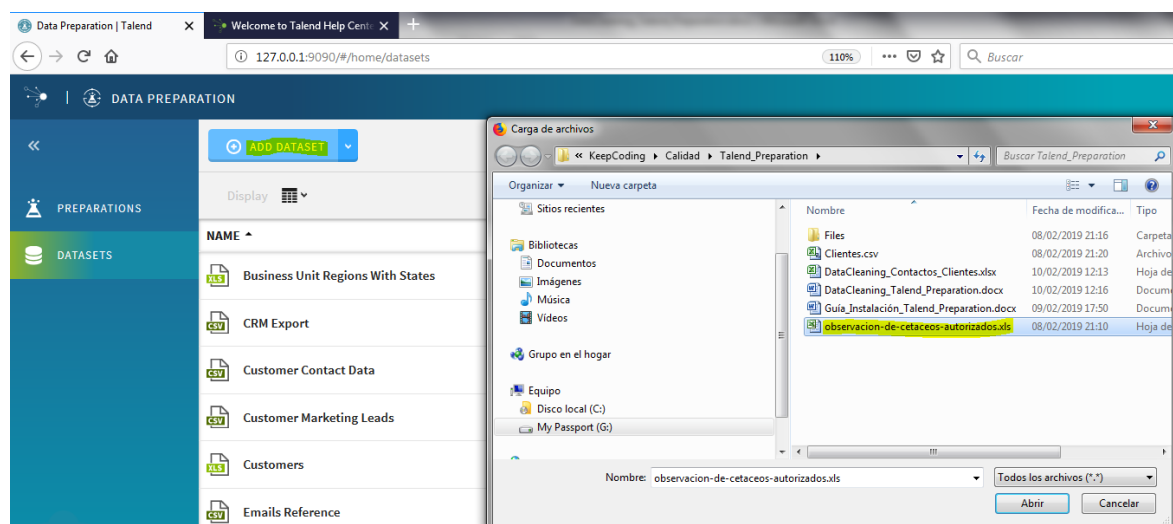




## Ejercicio 2.-Data Cleansing de Datos Abiertos

Vamos utilizar el fichero de datos abiertos del Gobierno de Canarias (<https://opendata.gobiernodecanarias.org/dataset/observacion-de-cetaceos-autorizados> ).

El dataset contiene observaciones de cetáceos por parte de empresas autorizadas en las costas de las islas. Son datos recogidos por diferentes empresas y por esta razón se registran datos con errores de calidad de datos. Se puede descargar en formato xls o csv.



### IMPORT EXCEL FILE

Dataset name:

Style sheet:

Content (100 first rows):

Isla	Municipio	Situación Adm...	Codigo de ide...	Signatura	Nombre come...	Dirección p
Fuerteventura	Antigua	En Funcionamie...	2017/01119	CT-35-2-0000002	ALBAKORA CAT	Otro Pue
Fuerteventura	Oliva (La)	En Funcionamie...	2016/02891	CT-35-2-0000001	BARRACUDA PERD...	Puerto d
Fuerteventura	Oliva (La)	En Funcionamie...	2017/03168	CT-35-2-0000004	CAMPUS STELAE	Otro Pue
Fuerteventura	Pájara	En Funcionamie...	2017/02695	CT-35-2-0000003	ODYSSEE TERCERO	Otro Pue
Fuerteventura	Pájara	En Funcionamie...	2018/00185	CT-35-2-0000005	PEDRA SARTAÑA	Otro Pue
Fuerteventura	Pájara	En Funcionamie...	2018/02678	CT-35-2-0000006	AITACARO	Otro Pue



Como somos una empresa de aventura de nueva creación en Canarias, queremos incluir entre nuestra oferta de servicios, la actividad de avistamiento de cetáceos ya que es un gran reclamo para los clientes. Para contactar con e mpresas que desarrollan esta actividad, necesitamos una lista de clientes con sus datos de contacto para incluirlos en nuestro sistema de gestión de clientes.

Utilizaremos la fuente de datos abiertos que ofrece el Gobierno Canario para crear la lista de contactos, pero antes tenemos que limpiar la fuente de datos porque contiene inconsistencia y muchos datos mezclados.

Vamos a utilizar las funciones de preparación de datos de Talend, con el objetivo de obtener el fichero de contactos de empresas que realizan la actividad de avistamiento de cetáceos.

Los pasos realizados para el Data Cleaning de la fuente de datos abiertos Cetáceos han sido: Primero revisamos a nivel de tabla si hay duplicados y registros con todos los campos vacíos, para ello hacemos clic en TABLE de la lista de sugerencias y aplicamos las funciones Delete empty row y Remove duplicate rows del grupo de funciones Data Cleansing.

Ahora nos centramos el campo Teléfonos/Internet, que contiene datos de teléfonos y correos electrónicos, vamos a intartar aplicar varias funciones de extracción y limpieza para conseguir campos de teléfonos consistentes y válidos.

El primer paso que haremos es prescindir del texto del principio “Telefonos :” creando una nueva columna sin los 10 primeros caracteres del campo, utilizaremos la función “Extract part of text”, escribiremos la primera palabra “Extrac” y buscaremos entre las opciones encontradas.

1 Extract parts of the text on column  
Teléfonos/Internet

☒ Create new column

From:  
From index

Beginning index:  
10

To:  
To end

Y borramos el campo original, seleccionando en el campo y elegimos “Delete column” del menú contextual.

Recordad que cualquier paso que demos que no nos dé el resultado esperado, podemos borrarlo.



Cambiamos el nombre de la nueva columna obtenida, y la volvemos a llamar Telefonos/Internet. Ahora quitamos espacios en blanco de la columna, utilizando la función “Remove trailing and leanding caracteres” de la lista de funciones.

Observamos que el separador que se utiliza para diferenciar la información, a veces es punto y como (;) y otras la coma (,) y otras el guión (-) la diferente información, vamos a cambiar el separador para que sea siempre el punto y coma y nos sirva para separar la información del campo, para ello utilizaremos la función “Search and Replace” en la misma columna, indicamos que busquemos la coma (,) y la remplazaremos por (;).

5 Search and replace on column  
Teléfonos/Internet

☐ Create new column

Search for:  
,,

Replace with:  
;

☐ Overwrite entire cell

SUBMIT

Haremos lo mismo con el separador de guion, utilizamos la función “Search and Replace” para cambiar el guión (-) por el (;).

El siguiente paso es utilizar el separador punto y coma para crear 2 nuevos campos desde el que tenemos, para ello utilizamos la función “Split the text in parts”

103/103

Teléfonos/Internet	Puerto
666 244 080;	Puerto Rico
928 851 448;	Puerto Rico
660 958 282; Correos: nauticapuertoja	Puerto Rico
928 540 133;	Puerto Rico
617 435 237;	Puerto Rico
928 150 010;655 991 903 ; 656 603 687	Puerto Rico
928 270 748;629.201.978; Correos: sor	Puerto Rico

Teléfonos/Internet

COLUMN ROW TABLE

Split

Split the text in parts...

Parts:  
2

Separator:  
;

SUBMIT


Borramos el campo Telefono/Internet, porque ya no nos sirve y así limpiamos el dataset. Seleccionamos el campo y con el menú contextual elegimos “Delete column”.



Ya tenemos el separado el primer teléfono, vamos a estandarizarlo, pero antes cambiamos el nombre de la columna, seleccionándola y botón contextual “Rename Column”, le llamaremos Teléfono\_1.


Eliminamos espacios en blanco en la columna Teléfono\_1, utilizando la función “Remove trailing and leanding characters”

Si observamos el patrón del teléfono tenemos algunos registros con el punto (.) y otros con espacio, como la mayoría son espacio, vamos a cambiar el punto (.) por el espacio, utilizando la función “Search and Replace”

11 Search and replace on column Teléfono 1 

☐ Create new column

Search for:  

 .

Replace with:

☐ Overwrite entire cell

**SUBMIT**

Desde la gráfica de patrones tenemos 2 registros sin espacio, los modificamos manualmente para incluir los espacios.

Y otros 2 registros con espacios cada 2 números, los modificamos manualmente para incluir los espacios cada tres números.

Hay un registro con valores inválidos, lo eliminamos. En la columna teléfono\_1, botón contextual, y elegimos la función “Clear invalid values”

Ahora que todos tienen el mismo patrón, vamos a dividir el campo en 3 campos para quitar los espacios y luego concatenarlos.

Utilizamos la función “Split the text in parts” para dividir el campo en 3 partes cuyo separador es el espacio en blanco.



### Teléfono 1

COLUMN ROW TABLE

split

Split the text in parts...

Parts:

3

Separator:

Space

SUBMIT

Ya tenemos divididas las tres partes del teléfono, ahora las concatenamos para obtener el teléfono sin espacios, utilizando la función “Concatenate with” y la opción “Other column”, primero con la segunda parte del teléfono obtenida del Split y luego con la tercera.

19 Concatenate with on column  
Teléfono 1\_split\_1

☐ Create new column

Prefix:

Use with:

Other column

Column:

Teléfono 1\_split\_3

Separator:

Add separator:

Both values not empty

Suffix:

Filters

Add a filter...

	xt	text	text	text	decimal	integer
1	CT-35-2-0000002	ALBAKORA CAT	Otro Puerto El Casti	666 244 080	666244080	
2	CT-35-2-0000001	BARRACUDA PERDOMO	Puerto de Corralejo,	928 851 448	928851448	
3	CT-35-2-0000004	CAMPUS STELAE	Otro Puerto de Corra			
4	CT-35-2-0000003	ODYSSEE TERCERO	Otro Puerto de Morro	660 958 282	660958282	
5	CT-35-2-0000005	PEDRA SARTAÑA	Otro Puerto de Morro	928 540 133	928540133	
6	CT-35-2-0000006	AITACARO	Otro Puerto de Morro	617 435 237	617435237	
7	CT-35-1-0000008	SPIRIT OF THE SEA	Puerto Rico, -Puerto	928 150 010	928150010	
8	CT-35-1-0000011	MULTIACUATIC	Puerto Rico, -Puerto	928 270 748	928270748	
9	CT-35-1-0000012	DOLPHIN'S CAT	Puerto Rico-Puerto R	928 243 685	928243685	
10	CT-35-1-0000013	SUPERCAT UNO	Puerto Rico-Puerto R	928 151 266	928151266	
11	CT-35-1-0000014	BAHIA CAT	Puerto Rico, -Puerto	928 355 981	928355981	
12	CT-35-1-0000015	FUNNY DAY UNO	Puerto Rico-Puerto R	928 243 685	928243685	
13	CT-35-1-0000010	BIRLOKE	Otro Puerto Rico, -	636 271 841	636271841	
14	CT-35-1-0000016	SALMON ONCE	Puerto Rico-Puerto R	928 243 708	928243708	
15	CT-35-1-0000017	SAGITTARIUS CAT	Puerto Rico - Atraque	638 439 037	638439037	
16	CT-35-1-0000018	EL TRUEQUE	Otro Puerto Rico - A	636 271 841	636271841	
17	CT-35-1-0000019	BLUE M	Otro Puerto de Mogán			

Eliminamos las 3 columnas Telefono\_1\_split\_2 y Telefono\_1\_splilt\_3 y Telefono\_1 porque ya no los necesitamos y renombramos Telefono\_1\_split\_1 por Telefono\_1, utilizando las funciones “Delete column” y “Rename column”

Ya hemos conseguido el número de Teléfono, ahora continuamos para obtener un segundo teléfono y email. Utilizaremos las mismas funciones que antes.



Sobre el campo obtenido por el primer Split, Teléfonos/Internet\_split\_2, quitamos espacios con la función “Remove trailing and leading characters”.

El siguiente paso es con “Split the text in parts” dividir el campo en 2 partes, como separador el punto y como (;)

The screenshot shows a data table with columns: Teléfonos/Internet\_split\_2, Puerto Base, and Titular. The table contains several rows of data. To the right, a configuration panel titled 'Teléfonos/Internet\_split\_2' is visible. It has tabs for COLUMN, ROW, and TABLE. The 'split' section is active, showing 'Split the text in parts...'. The 'Parts' field is set to 2, and the 'Separator' field is set to ;. A green SUBMIT button is at the bottom of the panel.

Teléfonos/Internet_split_2	Puerto Base	Titular
	Puerto de Corralejo	C.I.F. B35767375-
	Puerto de Corralejo	C.I.F. B76140128-
Correos: nauticapuertojandia@yahoo.es;	Puerto de Morro Jable	C.I.F. B35917657-
	Puerto de Morro Jable	C.I.F. B48403323-
	Puerto de Morro Jable	N.I.F. 33478426V-f
655 991 903 ; 656 603 687; Correos: spirit.res	Puerto Rico	C.I.F. B35522788-
629.201.978; Correos: sonia@multifinanzas.es;	Puerto Rico	C.I.F. B35838937-
609 161 299; Correos: canariasyachts@telefonica	Puerto Rico	C.I.F. B35125335-
687.518.544 ; 630.972.266; Correos: administra	Puerto Rico	C.I.F. B35422104-

Eliminamos el campo origen Teléfonos/Internet\_split\_2, con la función “Delete column” y renombramos la primera parte del que se ha obtenido, le llamamos Telefono\_2

Eliminamos los registros inválidos sobre la columna del Telefono\_2, con la función “Clear the cells with invalid values” del menú contextual.

Ahora, utilizamos la función “Search and Replace” para cambiar los registros con el separador punto “.” Por espacio en blanco.

### 31 Search and replace on column Teléfonos\_2

The screenshot shows the 'Search and Replace' configuration panel. It has a checkbox for 'Create new column' which is unchecked. The 'Search for:' field contains a period (.). The 'Replace with:' field is empty. At the bottom, there is a checkbox for 'Overwrite entire cell' which is also unchecked.



Ahora como con el hicimos con el campo “Teléfono 1”, separamos el campo en 3 partes, para posteriormente concatenar con la función “Split the text in parts”

The screenshot shows a table with 5 columns: 'Teléfonos\_2' (decimal), 'Teléfonos\_2\_spli...' (integer), 'Teléfonos\_2\_spli...' (integer), 'Teléfonos\_2\_spli...' (integer), and 'Teléfonos/' (integer). The data rows show phone numbers split into three parts. To the right, the 'Teléfonos\_2' configuration panel is open, showing the 'split' function with 'Parts' set to 3 and 'Separator' set to 'Space'. A 'SUBMIT' button is visible at the bottom of the panel.

Teléfonos_2	Teléfonos_2_spli...	Teléfonos_2_spli...	Teléfonos_2_spli...	Teléfonos/
655 991 903	655	991	903	655 60...
629 201 978	629	201	978	Correo...
609 161 299	609	161	299	Correo...

Eliminamos el campo “Teléfono\_2” con la función “Remove column” y renombramos la primera de las parte obtendidas de la división, con el nombre “Telefono\_2”

Y concatenamos utilizando la función “Concatenate with” con las columnas de la segunda y la tercera parte de la división.

The screenshot shows the 'Concatenate with' configuration panel on the left, with 'Create new column' checked, 'Prefix' empty, 'Use with' set to 'Other column', 'Column' set to 'Teléfonos\_2\_split\_3', 'Separator' empty, 'Add separator' set to 'Both values not empty', and 'Suffix' empty. To the right, the 'Filters' panel is open, showing a table with 4 columns: 'léfono 1' (integer), 'Telefono\_2' (integer), and 'Telé' (integer). The data rows show the concatenated phone numbers.

	léfono 1	Telefono_2	Telé
1	666244080		
2	928851448		
3			
4	660958282		
5	928540133		
6	617435237		
7	928150010	655991903	
8	928270748	629201978	
9	928243685	609161299	
10	928151266	687518544	
11	928355981	609507186	
12	928243685	609161299	
13	636271841	928770059	
14	928243708	649925918	
15	638439037	609359020	
16	636271841	928770059	
17			
18	609514466	928772222	
19	639727522		



Ya hemos obtenido el segundo teléfono. Ahora borramos las columnas que no necesitamos, Telefono\_2\_split\_2 y Telefono\_2\_split\_3, mediante la función “Delete column”

Nos centramos ahora en obtener el email de las empresas, lo primero renombramos la columna obtenida en el Split y le llamamos email, con la función “Rename column”

Observamos el patrón y seleccionamos los registros que no siguen el patrón “Aaaaaa: “, vamos a eliminar los registros filtrados, utilizando la función “Delete these filters rows”

The screenshot shows a data table with columns: email, email\_split\_1, email\_split\_2, and Puerto Base. A filter is applied to the 'email' column with the value '999 999 999; Aaaaa: 999 999 999;'. The right sidebar shows the 'email' column selected, and a suggestion to 'Delete these filtered rows' is highlighted.

Quitamos espacios sobre el email, con la función “Remove Trailing and leanding characters” y hacemos Split de 2 por el separador dos puntos (:)

The screenshot shows a data table with columns: léfono 1, Teléfono\_2, email, Puerto Base, and Titular. A filter is applied to the 'email' column. The right sidebar shows the 'email' column selected, and a suggestion to 'Split the text in parts...' is highlighted. The 'Split' function is configured with 'Parts: 2' and 'Separator: :'. A 'SUBMIT' button is visible.

Eliminamos la columna original email y la primera parte del Split.

Ahora volvemos hacer Split de 2 por el separador “;” para quedarnos sólo con la primera parte, que es la que contiene el email





The screenshot shows a Talend data integration workflow. On the left, a table named 'email\_split\_2' is displayed with columns 'Puerto Base' and 'Titular'. The table contains several rows of data, including contact information for various locations in Puerto Rico. On the right, the configuration for the 'email\_split\_2' component is shown. It is set to 'split' mode, and the configuration panel shows 'Split the text in parts...' with 'Parts' set to 2 and 'Separator' set to ';'. A 'SUBMIT' button is visible at the bottom of the configuration panel.

email_split_2	Puerto Base	Titular
	Puerto El Castillo	C.I.F. B358
	Puerto de Corralejo	C.I.F. B358
	Puerto de Corralejo	C.I.F. B761
	Puerto de Morro Jabl	C.I.F. B358
	Puerto de Morro Jabl	C.I.F. B484
	Puerto de Morro Jabl	N.I.F. 33478
sonia@multifinanzas.es; Faxes: 928 279	Puerto Rico	C.I.F. B358
canariasyachts@telefonica.net; Faxes: 9	Puerto Rico	C.I.F. B358
administracion@grupotomassosa.com; Pági	Puerto Rico	C.I.F. B358

Borramos las columnas que no nos hacen falta y renombramos la columna que se queda con el email, le llamamos “email”, además cambiamos el tipo semántico a Tipo email.

Eliminamos del data set los registros que tengan vacíos los campos de Telefono\_1, Telefono\_2 y email, porque no es necesario su información al no tener ningún campo de contacto informado.

Ahora si quisiéramos podemos enriquecer también la columna de CIF/NIF para separar el CIF del Titular o exportar los datos que ya hemos limpiado porque hemos conseguido los datos de contacto.

El fichero XLS obtenido de los pasos de la preparación con Talend se llama “observacion-de-cetaceos-autorizados Preparation.xlsx”