# **Tutorial Talend Open Studio for Big Data**

Cómo crear un proceso para automatizar la normalización tablas de indicadores del banco mundial

#### 1) Descargar los datos fuente

Descargar del Banco Mundial las tablas de todos los <u>indicadores</u> que deseemos procesar y comparar como archivos Excel.

Para el ejercicio usaremos:

Población de refugiados por país o territorio de asilo

Población de refugiados por país o territorio de origen

#### 2) Abrir Talend y crear un nuevo job.



En el panel de la izquierda: Repositorio, en la rama de Job Designs podemos dar clic derecho para que se despliegue un menú de opciones. La primera opción es para crear un nuevo Job.

El nuevo Job necesita al menos un nombre que debe cumplir con los requisitos para nombrar variables en el Lenguaje Java.

• Alfanumérico pero NO debe comenzar con números.

- No debe contener espacios en blanco
- Se puede usar el caracter \_
- Distingue Mayúsculas y minúsculas

📵 New Job			
New Job			
🚷 Name is	empty.		
Name			
Purpose			
Description			 ^
			~
Author	user@talend.com		
Locker			
Version	0.1		Mm
Status			Ŷ
Path			Select
		Finish	Cancel

### 3) Extraer los datos.

Buscamos y arrastramos de la paleta al área de trabajo el componente para leer los datos de un archivo Excel: tFileInputExcel

Cuando nos colocamos encima del componente, en el tab Componente en la parte inferior se mostrarán los parámetros de configuración del componente.

Ahí debemos definir:

- 1) La ruta del archivo Excel a extraer (File name)
- 2) El nombre de la hoja a extraer (Sheet list)
- 3) La posición de Inicio del dataset (Header)
- 4) La última columna a leer (Last column)



		tFileInputExcel_1	
ner Code			
ob(Turorial 0.1)	Contexts(Tu	orial) 🕫 Component 🗴 🕩 Run (Job Turorial)	⊕ ⊕
tFileInputEx	cel_1		
settings	Property Type	Built-In 🗸 🖬	
nced settings	Read excel2007	file format(xlsx)	
mic settings	File name/Stream	"J:/Talleres/journocoders/Refugiados por pais de origen.xls"	*
	All sheets		
mentation	Sheet list	Sheet (name or position) Use Regex	
		"Data"	
	Header	4 Footer 0 Limit	
	Affect each she	et(header&footer)	
	Die on error		
	First column	1 Last column 60	
	Schema	Built-In 🗸 Edit schema 🐃	

Con el botón "Edit Schema" se despliega una ventana donde podemos definir los campos de la tabla y sus metadatos:

countryName   String   Image: String     countryCode   String   Image: String     indicatorName   String   Image: String     indicatorCode   String   Image: String     a1960   Double   Image: String     a1961   Double   Image: String     a1962   Double   Image: String     a1963   Double   Image: String     a1964   Double   Image: String     a1965   Double   Image: String     a1966   Double   Image: String	Column	Key	Туре	✓ N	Date Pattern (Ctr	Length	Precision	Default	Comment
countryCode   String     indicatorName   String     indicatorCode   String     a1960   Double     a1961   Double     a1962   Double     a1963   Double     a1964   Double     a1965   Double	countryName		String			-			
indicatorName   String   Image: String     indicatorCode   String   Image: String     a1960   Double   Image: String     a1961   Double   Image: String     a1962   Double   Image: String     a1963   Double   Image: String     a1964   Double   Image: String     a1965   Double   Image: String     a1966   Double   Image: String	countryCode		String	$\checkmark$					
indicatorCode String Image: Constraint of the string of t	indicatorName		String	$\checkmark$					
a 1960 Double   a 1961 Double   a 1962 Double   Double Image: Comparison of the state of the	indicatorCode		String	$\checkmark$					
a1961 Double Image: Comparison of the com	a1960		Double	$\checkmark$					
a 1962 Double Image: Comparison of the co	a1961		Double	$\checkmark$					
a 1963 Double Image: Comparison of the co	a1962		Double	$\checkmark$					
a 1964 Double Image: Comparison of the co	a1963		Double	$\checkmark$					
a1965 Double 2	a1964		Double	$\checkmark$					
	a1965		Double	$\checkmark$					
	s1066		Double						

## 4) Transformar los datos

Luego, debemos arrastrar otro componente que se llama TFlowTolterate en la categoría de **Orchestration**,



Ahora vamos a unir esos dos componentes por medio del flujo de datos principal.

Para esto pinchamos con el botón derecho sobre el componente tFileInputExcel y seleccionamos el flujo de datos principal, como muestra la imagen:



Sin tocar antes nada más, ahora pinchamos el componente tFlowToIterate, de esta manera los componentes quedan ligados:



Y lo que estamos diciendo es: el componente tFileInputExcel va a extraer la tabla del Excel y va a pasar los datos al componente tFlowTolterate.

El componente tFlowTolterate sirve para aplicar un proceso a CADA UNO de los registros en el conjunto de datos que recibe.

A continuación, arrastramos el componente tLoop al área de trabajo y lo conectamos al tFlowTolterate por medio del link **iterate**.

tFileInputExcel_1	row1 (Main)							
Contexts(Turorial)								
Loop Type For While								
From	1960	-						
То	2015	-						
Step	1	-						
Values are incre	·							

Este componente nos va a servir para aplicar un conjunto de instrucciones varias veces según un valor que se utiliza como "índice", el cual va a comenzar con el valor 1960, que es el primer año de la data hasta 2015.

En otras palabras vamos a ejecutar un proceso para cada registro del Excel y para cada año desde 1960 hasta 2015

Seguido, vamos a arrastrar y a ligar un componente llamado tFixedFlowInput, que sirve para crear dinámicamente un flujo de datos según las columnas que definamos y el valor que les asignemos.

Primero vamos a definir el esquema para este componente (Clic sobre "Edit Schema")

Schema of tFixedFlowIng	put_1							×
tFixedFlowInput_1								
Column	Key	Туре	✓ N	Date Pattern (Ct	Length	Precision	Default	Com
countryName		String	$\checkmark$					
countryCode		String	$\checkmark$					
ano		Integer	$\checkmark$					
refOrigen		Double						
•								
🕂 🗙 🗘 🕀 🗎		Q 6	1					
						ОК	Cance	

	tFileInputExcel_1	row1 (Main)	tFlowTolterate_1	Iterate	D D tLoop_1	lterate	tFixedFlowInput_1		
al 0.1)	Contexts(Tu	rorial) 🔭 Componen	t 🗙 🕩 Run (Job Turor	ial)					
lowin	iput_1								
	Schema	Built-In 🗸 Edit schema							
ngs	Number of rows	ws 1							
gs	Mode Use Single Table	le							
n	Values	Column	Value						
		countryName	((String)globalMap.	get("row1.counti	ryName"))				
		countryCode	((String)globalMap.	get("row1.counti	ryCode"))				
		ano	((Integer)globalMap	o.get("tLoop_1_C	URRENT_VALUE"))				
		refOrigen	((Double)globalMap	p.get("row1.a"+ (	(Integer)globalMap.get("t	Loop_1_CURREI	NT_VALUE"))))		
	O Use Inline Tabl	e							
	O Use Inline Cont	tent(delimited file)							

Luego, establecemos el valor que va a tener cada columna de este nuevo set:

Los valores vienen de los 2 componentes anteriores, son autocompletados, basta con digitar "TFlow" + [TAB] y nos salen los valores actuales que maneja este componente

Para el valor correspondiente al año vamos a utilizar el valor actual del Loop para obtener el valor de la columna deseado.

### 5) Y Finalmente Cargar de los datos

Ahora solo es cuestión de escoger el formato de salida de los datos ya transformados, por ejemplo si queremos generar un csv hay un componente para ésto: tFileOutputDelimited, al cual le enviamos el flujo de datos principal (Main), le definimos la ruta del archivo a generar y los demás parámetros de configuración según muestra la imagen:

tEilelo		
tritein	tFlowTolterate_1 tLoop_1 tFixedFlowInput_1 tFileOutputDelimited_1	
rsPIB 0.1) 📴 Cont	texts(journocodersPIB) 🕫 Component × 🕨 Run (Job journocodersPIB) $igodot$	ŧ
Delimited_1		_
Property Type	Built-In 🗸	
Use Output Stre	am	
File Name	"J:/Talleres/journocoders/CO2 Emisiones Normalizado.csv"	••
Row Separator	"\n" Field Separator ";"	_
Append Inc	clude Header	
Schema	Built-In 🔻 Edit schema 📨 Sync columns	

Y ya estamos listos para poner a correr el Job, lo hacemos desde el tab "Run" donde hay un botón de "Run"

Designer Code				
Job(journocoders	;PIB 0.1)  🛐	Contexts(journocodersPIB)	😵 Component	▶ Run (Job journocodersPIB) ×
Job journocoders	sPIB			
Basic Run Debug Run	Execution	Run 🗏 Kill	Clear	
Advanced settings Target Exec				
Memory Run				