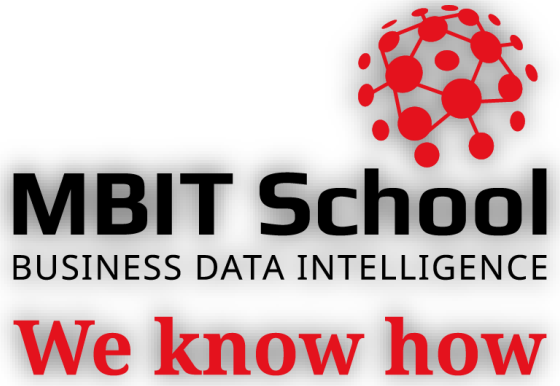


M01-DC03 ETL PENTAHO

DC03 TAREA EVALUABLE PENTAHO

Programa: **Máster Executive en Big Data, Cloud & Analytics**

Periodo académico: **2019 – 2020**

Autor/es: **CARLOS ALFONSEL JAÉN**

1. ENUNCIADO

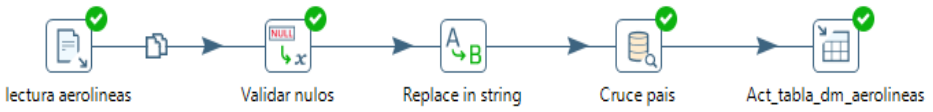
La tarea evaluable consiste en realizar la transformación para la carga de la dimensión aerolíneas que no hicimos en clase del caso práctico Pentaho.

El objetivo es realizar los pasos indicados en el documento 'Taller Pentaho aeropuertos' para los datos de aerolíneas, sobre los pasos aquí indicados se deja como opcional añadir pasos adicionales sobre los campos del dataset (ordenaciones, cálculo de campos nuevos, otras transformaciones) utilizando alguno de los pasos estándar disponibles en Pentaho (a elegir por el alumno). Una opción es obtener un campo adicional, aunque este no se inserte en la tabla destino en el paso de 'salida a tabla'. La transformación deberá ser integrada en el job realizado en clase

El entregable es el fichero .ktr correspondiente a la transformación, el fichero .kjb con el job que integra la transformación más los documentos explicativos que el alumno considere.

2. SOLUCIÓN

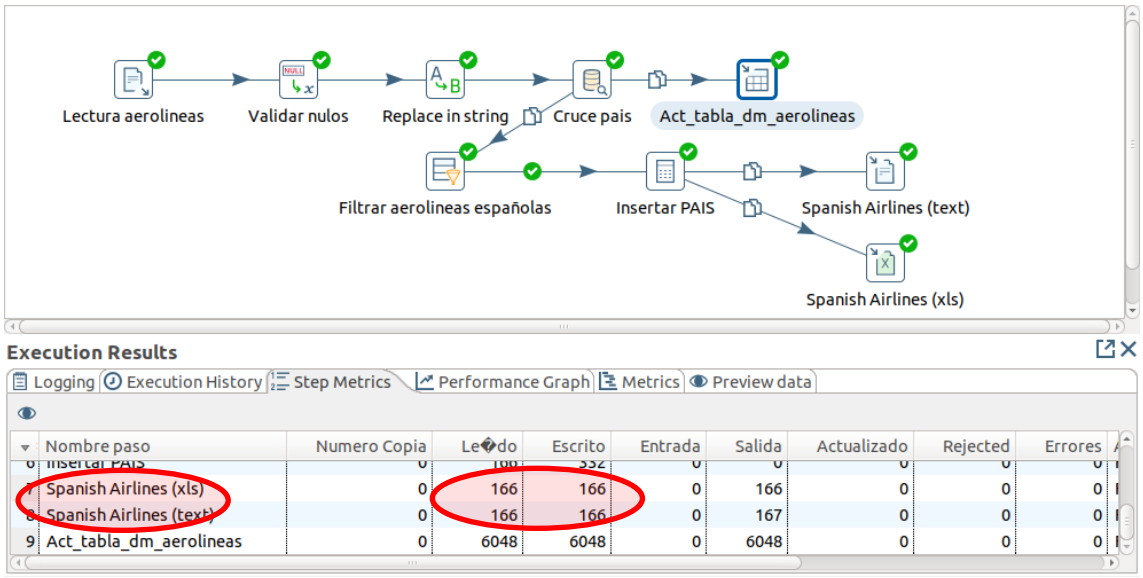
La transformación básica planteada en el enunciado es la siguiente:



Esta transformación se ha enriquecido con los siguientes pasos:

- Un filtrado previo de las aerolíneas con bandera española (filtro pais_aerolinea = Spain).
- Un campo nuevo calculado en el que se añade el string pais_aerolinea con todas sus letras en MAYÚSCULAS.
- El resultado del filtrado más el campo nuevo añadido se vuelca en un archivo de texto Spanish_Airlines.txt y en una hoja Excel Spanish_Airlines.xls.

Así que la nueva transformación tr_aerolineas queda de la siguiente manera:

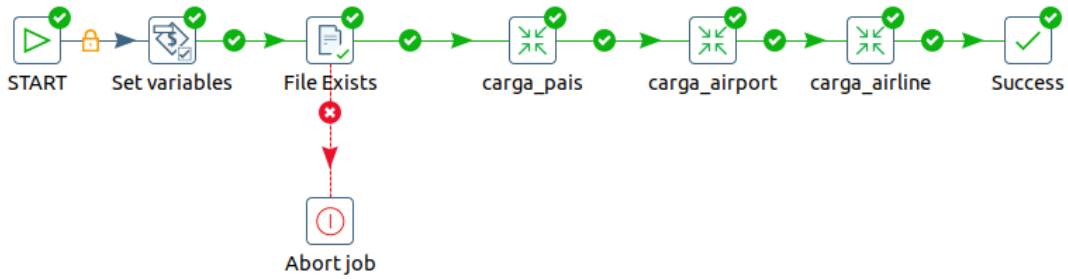


Execution Results

Logging | Execution History | Step Metrics | Performance Graph | Metrics | Preview data

	Nombre paso	Numero Copia	Le?do	Escrito	Entrada	Salida	Actualizado	Rejected	Errores
7	Spanish Airlines (xls)	0	166	166	0	166	0	0	0
8	Spanish Airlines (text)	0	166	166	0	167	0	0	0
9	Act_tabla_dm_aerolineas	0	6048	6048	0	6048	0	0	0

Asimismo, se ha ejecutado el job_aero incluyendo esta nueva transformación:



pentaho_job.log: Bloc de notas

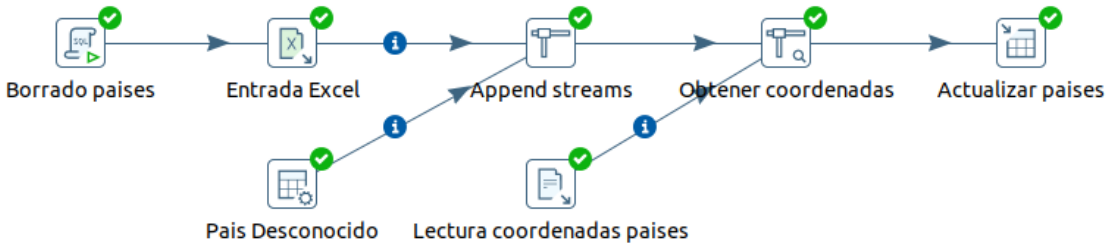
Archivo Edición Formato Ver Ayuda

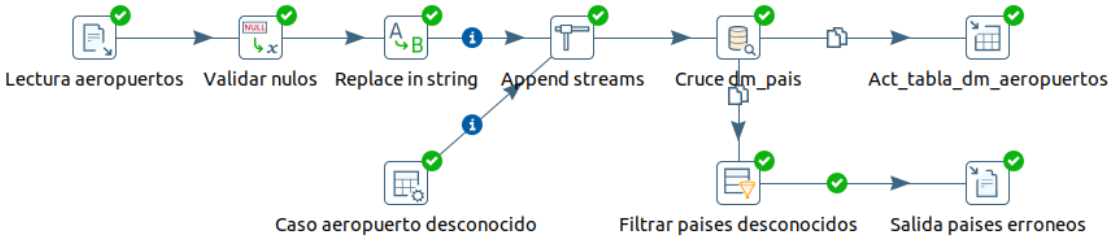
```
2020/01/19 20:58:32 - carga_airline - Using run configuration [Pentaho local]
2020/01/19 20:58:32 - carga_airline - Using legacy execution engine
2020/01/19 20:58:32 - tr_aerolineas - Iniciado despacho de la transformaci♦n [tr_aerolineas]
Sun Jan 19 20:58:32 CET 2020 WARN: Establishing SSL connection without server's identity verification is not recommended. A
Sun Jan 19 20:58:32 CET 2020 WARN: Establishing SSL connection without server's identity verification is not recommended. A
2020/01/19 20:58:32 - Act_tabla_dm_aerolineas.0 - Connected to database [curso_bbdd] (commit=1000)
2020/01/19 20:58:33 - Lectura aerolineas.0 - Procesamiento finalizado (I=6048, O=0, R=0, W=6048, U=0, E=0)
2020/01/19 20:58:33 - Validar nulos.0 - Procesamiento finalizado (I=0, O=0, R=6048, W=6048, U=0, E=0)
2020/01/19 20:58:33 - Replace in string.0 - Procesamiento finalizado (I=0, O=0, R=6048, W=6048, U=0, E=0)
2020/01/19 20:58:36 - Cruce pais.0 - Procesamiento finalizado (I=5872, O=0, R=6048, W=12096, U=0, E=0)
2020/01/19 20:58:36 - Filtrar aerolineas españolas.0 - Procesamiento finalizado (I=0, O=0, R=6048, W=166, U=0, E=0)
2020/01/19 20:58:36 - Insertar PAIS.0 - Procesamiento finalizado (I=0, O=0, R=166, W=332, U=0, E=0)
2020/01/19 20:58:36 - Spanish Airlines (text).0 - Procesamiento finalizado (I=0, O=167, R=166, W=166, U=0, E=0)
2020/01/19 20:58:36 - Spanish Airlines (xls).0 - Procesamiento finalizado (I=0, O=166, R=166, W=166, U=0, E=0)
2020/01/19 20:58:37 - Act_tabla_dm_aerolineas.0 - Procesamiento finalizado (I=0, O=6048, R=6048, W=6048, U=0, E=0)
2020/01/19 20:58:37 - job_aero - Starting entry [Success]
2020/01/19 20:58:37 - job_aero - Finished job entry [Success] (result=[true])
2020/01/19 20:58:37 - job_aero - Finished job entry [carga_airline] (result=[true])
2020/01/19 20:58:37 - job_aero - Finished job entry [carga_airport] (result=[true])
2020/01/19 20:58:37 - job_aero - Finished job entry [carga_pais] (result=[true])
2020/01/19 20:58:37 - job_aero - Finished job entry [File Exists] (result=[true])
2020/01/19 20:58:37 - job_aero - Finished job entry [Set variables] (result=[true])
2020/01/19 20:58:37 - job_aero - Job execution finished
2020/01/19 20:58:37 - Kitchen - Finished!
2020/01/19 20:58:37 - Kitchen - Start=2020/01/19 20:55:59.510, Stop=2020/01/19 20:58:37.155
2020/01/19 20:58:37 - Kitchen - Processing ended after 2 minutes and 37 seconds (157 seconds total).
```

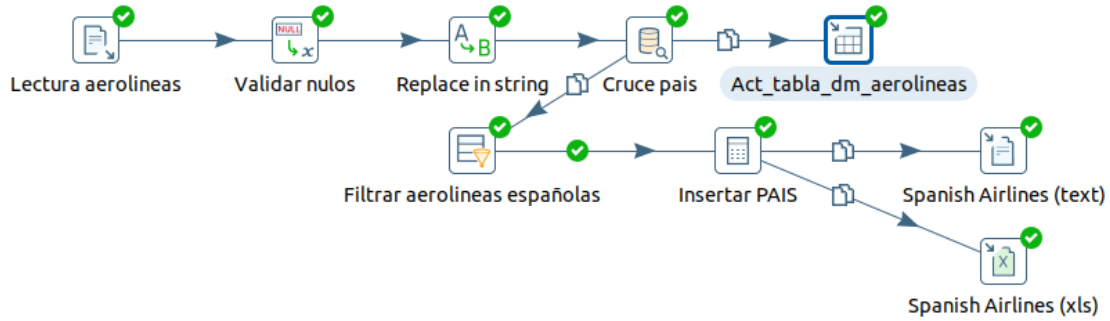
UNIX (LF) | Línea 1, columna 1 | 100%

Siendo el resultado del mismo exitoso, como puede verse en el pantallazo anterior del log obtenido mediante el comando `sh kitchen.sh -file=/home/mbit/taller_pentaho/soluciones/job_aero.kjb -level=Basic > /home/mbit/taller_pentaho/logs/pentaho_job.log` ejecutado por consola.

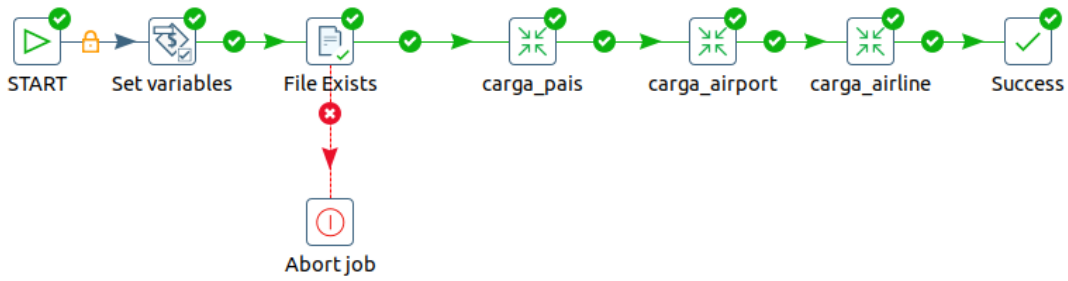
En resumen, las tres transformaciones (tr_pais, tr_aeropuertos y tr_aerolineas) quedan de la siguiente manera:







Mientras que el job_aero:



En los entregables de la práctica se incluyen los siguientes ficheros:

- Carpeta soluciones:
  - job_aero.kjb
  - tr_aerolineas.ktr
  - tr_aeropuertos.ktr
  - tr_pais.ktr
- Carpeta logs:
  - Pentaho_job.log
- Carpeta ficheros:
  - paises_error.txt
  - spanish_airlines.txt
  - spanish_airlines.xls