

Cas pràctic: Magatzem de dades per l'anàlisi de la conciliació laboral i familiar o *work-life balance*

PR2 – Càrrega de dades

Índex

Presentació.....	2
Descripció.....	3
Guia de mostra	4
Criteris d'avaluació	5
1. Identificació dels processos ETL.....	6
2. Disseny i desenvolupament dels processos ETL.....	8
2.1. Creació de taules.....	8
2.1.1. Bloc IN	8
3. Implementació dels treballs amb processos ETL.....	13
Format i data de lliurament.....	14

Presentació

A partir de la solució oficial de la primera pràctica (PR1), l'estudiantat ha de dissenyar, implementar i executar els processos d'extracció, transformació i càrrega de les dades (ETL) relatives a indicadors destinats a l'anàlisi de la conciliació familiar a partir de les fonts de dades proporcionades.

Així doncs, aquesta activitat té com a objectiu identificar i desenvolupar els processos de càrrega del magatzem de dades (DW) i fer-la efectiva.

Per dur-la a terme, s'ha publicat, al costat d'aquest enunciat, el fitxer **DDL_STG+DW.sql** amb els *scripts* de creació de les taules de l'*staging area* (STG) i del magatzem de dades.

Descripció

Si ens centrem en els subobjectius, aquesta segona part del cas pràctic consisteix en el següent:

- Identificar els processos d'ETL cap al magatzem de dades.
- Dissenyar i desenvolupar els processos d'ETL mitjançant les eines de disseny proporcionades.
- Implementar amb els treballs (*jobs*) els processos d'ETL perquè la seva càrrega planificada sigui efectiva.

A més del document amb la solució de la PR2 que s'ha de lliurar, també es tindrà en consideració la implementació sobre la màquina virtual proporcionada en el curs.

En resum, el document de la solució de la PR2 ha d'incloure els aspectes següents:

- Descripció de totes les accions que s'han portat a terme.
- Captures de pantalla on es mostrin totes les parts significatives dels processos ETL, les seves característiques i la seva explicació corresponent.
- Captures de pantalla que demostrin la correcta execució dels processos ETL i el temps d'execució.
- Captures de pantalla que demostrin la correcta càrrega de les dades (carregades en la base de dades).

Guia de mostra

Amb la finalitat d'ajudar a aconseguir els objectius plantejats de la PR2, s'ha desenvolupat aquesta guia de mostra per ser utilitzada com a exemple de com realitzar alguna de les tasques descrites anteriorment, és a dir, el disseny i el desenvolupament dels processos ETL i la càrrega efectiva al magatzem de dades.

Criteris d'avaluació

S'han de justificar les decisions preses acompanyant-se de les captures de pantalla, amb els passos més significatius que ho acreditin.

1) Identificació dels processos ETL	20 %
a) Identificació i descripció dels processos ETL d'ORIGEN a STAGE	10 %
b) Identificació i descripció dels processos ETL d'STAGE a DW	10 %
2) Disseny i desenvolupament dels processos ETL	60 %
a) Transformacions Bloc IN	20 %
b) Transformacions Bloc TR_DIM	20 %
c) Transformacions Bloc TR_FACT	20 %
3) Implementació dels treballs dels processos ETL	20 %
a) Jobs Bloc IN	5 %
b) Jobs Bloc TR_DIM	5 %
c) Jobs Bloc TR_FACT	5 %
d) Procés complet (utilitza els Jobs IN i TR)	5 %

1. Identificació dels processos ETL

A l'hora de dissenyar els processos de càrrega d'una base de dades analítica no hi ha una estratègia única. És habitual estructurar els processos ETL sobre la base de les entitats de dades que s'han d'actualitzar, ja que hi ha diferències conceptuais en l'actualització d'una dimensió respecte a la d'una taula de fets. La divisió del procés de càrrega inicial en diferents blocs d'actualització facilitarà el disseny d'un ordre d'execució i la gestió de les dependències. Cadascun d'aquests blocs d'actualització es dividirà en les corresponents etapes d'ETL.

S'identifiquen els dos blocs següents:

- **Bloc IN:** processos de càrrega de les dades des de les fonts fins a les taules intermèdies en l'àrea de maniobres (STG). Aquests processos es distingeixen pel prefix IN_ en el nom.
- **Bloc TR:** processos de transformació per carregar les dades des de les taules intermèdies fins al magatzem, segons el model multidimensional dissenyat. Així doncs, els processos ETL de transformació per carregar les dimensions són diferents d'aquells que es porten a terme per carregar les taules de fets. Aquests processos es distingeixen amb el prefix TR_ en el nom.

A continuació s'identifiquen **alguns dels processos** que formen part de cadascun dels blocs d'actualització:

Bloc IN (de les fonts a les taules intermèdies)

Nom de l'ETL	Descripció	Orígens de les dades	Taula de destinació (stage)
IN_ESTAT_AGE	Càrrega de la informació relativa al tram d'edat	ESTAT_AGE_en.tsv	STG_ESTAT_AGE
...

Bloc TR (de les taules intermèdies al magatzem)

El bloc TR de processos ETL per poblar el model multidimensional del magatzem té dues parts diferenciades. D'una banda, els processos de càrrega i transformació de les dimensions i, per un altre, els de les taules de fets. L'ordre d'execució és important perquè la càrrega de dades sigui la correcta. Les dimensions es carregaran primer i, després, les taules de fets perquè no hi hagi errors durant la càrrega.

D'una banda, alguns dels processos del bloc TR de càrrega i transformació de les dimensions són els següents:

Nom de l'ETL	Descripció	Taula d'origen	Taula de destinació (dimensions)
TR_DIM_AGE	Càrrega de la informació relativa al tram d'edat	STG_ESTAT_AGE	DIM_AGE
...

Els processos de càrrega de les dimensions són molt semblants en alguns casos. En el lliurament s'han d'identificar tots, encara que es podrien exposar de manera conjunta en alguns dels casos. Per a això, s'ha d'exposar clarament quins processos de càrrega són idèntics i explicar detalladament tots els passos per a un d'ells solament. Qualsevol altre procés de càrrega que sigui diferent s'haurà d'explicar de manera íntegra.

D'altra banda, s'han d'identificar els processos de càrrega dels fets i les seves taules d'origen i destinació.

Nom de l'ETL	Descripció	Taula d'origen	Taula de destinació (dimensions)
TR_FAC_PCT_EMPLOYEES_HOME	Càrrega de la informació relativa al percentatge de persones que treballen a casa	STG_PCT_EMPL OYEES_HOME	FACT_PCT_EMPL OYEES_HOME
...

En aquest punt, l'estudiantat haurà de completar la identificació dels processos de cadascun dels blocs (IN i TR) que desenvoluparà per carregar les dimensions, així com les taules de fets del model multidimensional del magatzem de dades.

2. Disseny i desenvolupament dels processos ETL

En aquest apartat s'han de dissenyar els processos de càrrega identificats en el punt anterior amb l'eina de disseny proporcionada. En aquest cas és *Pentaho Data Integration* (PDI).

2.1. Creació de taules

El primer pas per a la implementació dels processos ETL consisteix en la creació de les taules. Això es durà a terme una única vegada mitjançant scripts i sobre la base de dades proporcionada (en aquest cas, SQL Server). S'hauran de crear les taules intermèdies i les taules del model dimensional de la solució oficial, és a dir, les dimensions i les taules de fets. Per fer-ho, s'han d'utilitzar els *scripts* facilitats juntament amb l'enunciat de la PR2.

Una vegada tingueu implementat el model físic del magatzem, el següent pas que cal portar a terme és el disseny dels processos ETL de cadascun dels blocs (IN i TR). Aquests processos permetran poblar les taules de l'àrea intermèdia (STG), les de dimensions i les de fets del magatzem de dades que heu dissenyat.

2.1.1. Bloc IN

Transformació d'IN_ESTAT_AGE

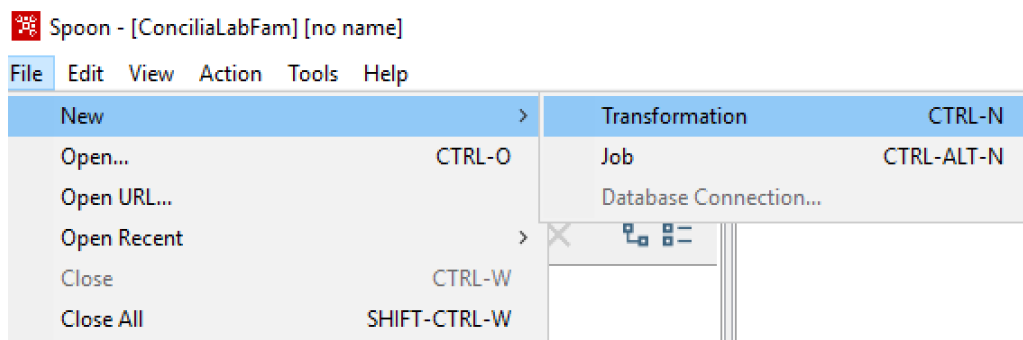
A continuació es descriu una part del desenvolupament de la transformació d'IN_ESTAT_AGE mitjançant Spoon. L'objectiu és carregar un dels orígens de les dades identificades, ESTAT_AGE_en.tsv, en la taula STG_ESTAT_AGE de l'àrea intermèdia (STG). La taula intermèdia haurà d'haver estat creada amb anterioritat en la base de dades analítica, i el seu *script* s'haurà escrit en l'apartat «creació».

La transformació d'IN_ESTAT_AGE conté les etapes següents:

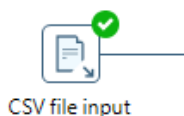
- Lectura del fitxer .tsv.
- Càrrega a la taula intermèdia STG_ESTAT_AGE.

A continuació es detallen les diferents etapes que s'implementaran per poder fer la càrrega de dades.

En primer lloc, es crea una nova transformació.



En aquesta, com a primer pas, s'utilitzarà l'entrada del fitxer TSV. Per dur-ho a terme, s'utilitzarà el tipus CSV Input. En aquest pas, s'haurà d'indicar el fitxer des d'on s'extrauran les dades. Per fer-ho, s'utilitzarà la variable d'entorn DIR_IN creada per indicar de manera única la ubicació dels fitxers i s'indicarà el tipus de motor que haurà d'usar.



Per fer correctament la càrrega, s'hauran d'indicar els paràmetres correctament i destacar, en aquest cas:

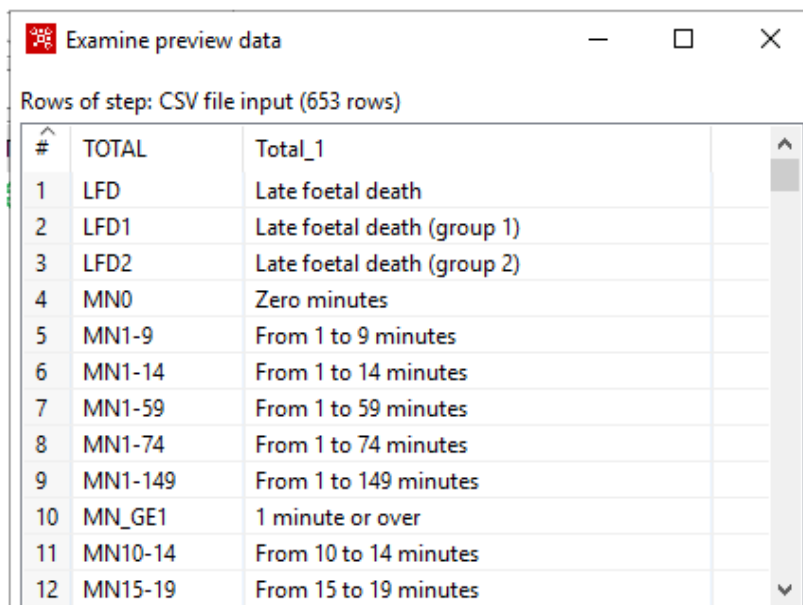
- Indicar com a delimitador TAB. Per a això, s'haurà de prémer el botó Insert TAB i es generarà automàticament un valor en blanc com es mostra en la imatge.
- Deseleccionar 'Lazy conversion?'.
- Seleccionar 'Header row present?'.

The screenshot shows the 'CSV file input' configuration window. The 'Step name' is 'CSV file input'. The 'Filename' is set to '\${DIR_IN}\ESTAT_AGE_en.tsv'. The 'Delimiter' is set to 'Insert TAB'. The 'Enclosure' is set to '"'. The 'NIO buffer size' is set to '50000'. The 'Lazy conversion?' checkbox is unchecked. The 'Header row present?' checkbox is checked. The 'Add filename to result' checkbox is unchecked. The 'The row number field name (optional)' field is empty. The 'Running in parallel?' checkbox is unchecked. The 'New line possible in fields?' checkbox is unchecked. The 'Format' is set to 'mixed'. The 'File encoding' is set to 'UTF-8'. Below the configuration fields, there is a table showing the output schema:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	TOTAL	String		13		\$,	.	none
2	Total	String		46		\$,	.	none

En aquest pas, també s'hauran d'indicar els camps que es tractaran amb el botó Get Fields i completar-ne la definició. Cal especificar, on es consideri necessari, la precisió i la longitud dels camps.

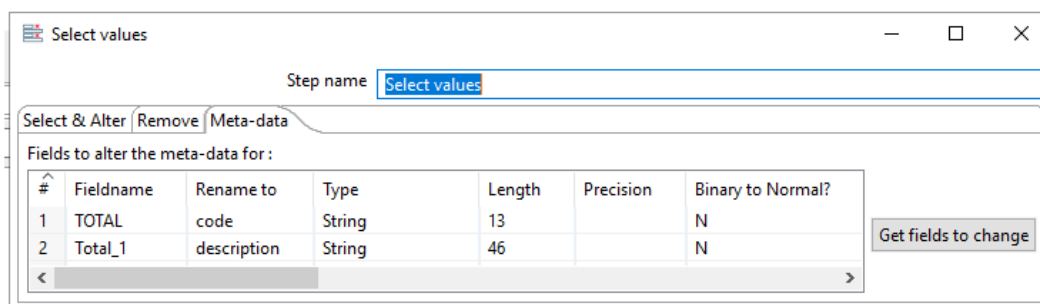
Es podrà fer una visualització prèvia de les dades, que es carregaran amb el botó Preview.



Rows of step: CSV file input (653 rows)

#	TOTAL	Total_1
1	LFD	Late foetal death
2	LFD1	Late foetal death (group 1)
3	LFD2	Late foetal death (group 2)
4	MN0	Zero minutes
5	MN1-9	From 1 to 9 minutes
6	MN1-14	From 1 to 14 minutes
7	MN1-59	From 1 to 59 minutes
8	MN1-74	From 1 to 74 minutes
9	MN1-149	From 1 to 149 minutes
10	MN_GE1	1 minute or over
11	MN10-14	From 10 to 14 minutes
12	MN15-19	From 15 to 19 minutes

Finalment, es podrà fer la transformació del tipus de dades amb el component Select Values, informant tots els camps de tipus *String*.



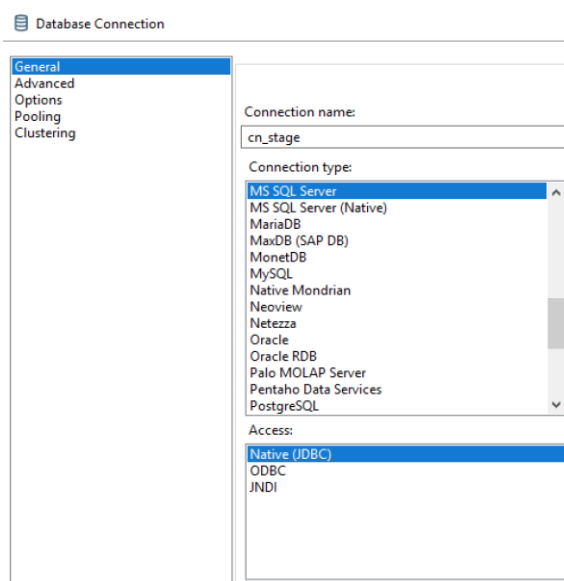
Step name:

Select & Alter Remove Meta-data

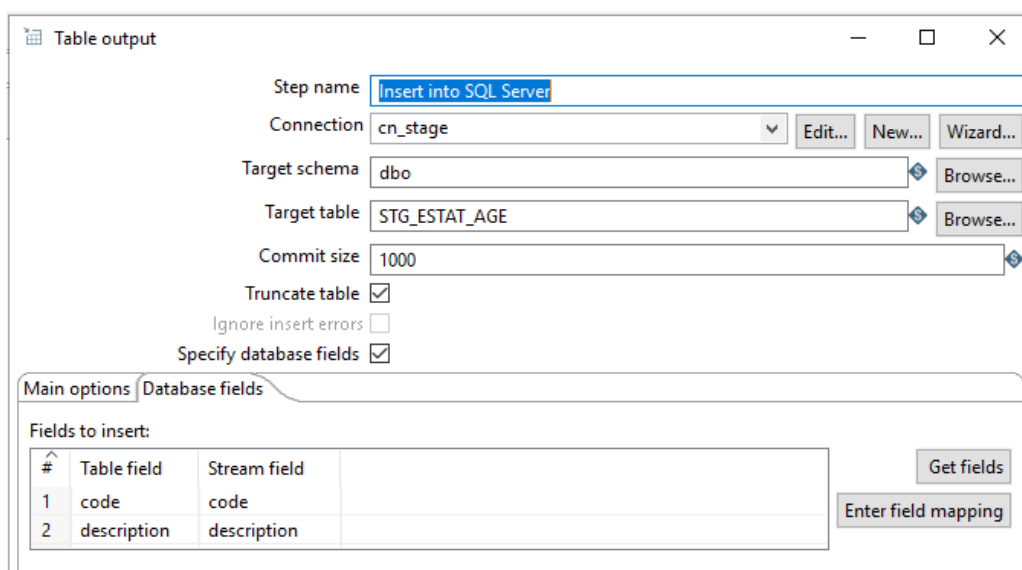
Fields to alter the meta-data for:

#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?
1	TOTAL	code	String	13		N
2	Total_1	description	String	46		N

Finalment, es carregaran les dades en la taula intermèdia de l'*stage* utilitzant el pas Table Output. Per a aquest pas, serà necessari especificar la connexió de la base de dades, que es podrà generar dins del mateix component prement el botó New i indicant que es vol connectar a Microsoft SQL Server i utilitzar l'accés natiu amb les vostres credencials. També es podran utilitzar les variables d'entorn creades, com ara la cadena de connexió.



El pas de carregar les dades a la taula intermèdia de l'*stage* es configurarà tal com s'indica en el menú principal que es mostra a continuació. És important que es tingui la taula d'*STG* creada per poder inserir les dades preprocessades.



Com s'aprecia en la imatge, per deixar la transformació preparada per a possibles reprocessaments, és necessari fer un esborrament previ per actualitzar les dades. Per a això, s'activarà el *check* Truncate Table, situat en els camps de la base de dades.

El procés de la transformació completa és el següent:



I es podrà comprovar en la base de dades el resultat del procés dut a terme:

```

SELECT [code]
      , [description]
FROM [dbo].[STG_ESTAT_AGE]
  
```

100 %

Results Messages

	code	description
1	LFD	Late foetal death
2	LFD1	Late foetal death (group 1)
3	LFD2	Late foetal death (group 2)
4	MN0	Zero minutes
5	MN1-9	From 1 to 9 minutes
6	MN1-14	From 1 to 14 minutes
7	MN1-59	From 1 to 59 minutes
8	MN1-74	From 1 to 74 minutes
9	MN1-149	From 1 to 149 minutes
10	MN_GE1	1 minute or over

Per al lliurament de la PR2, l'estudiantat haurà de dissenyar tots els processos ETL de cadascun dels blocs (IN i TR).

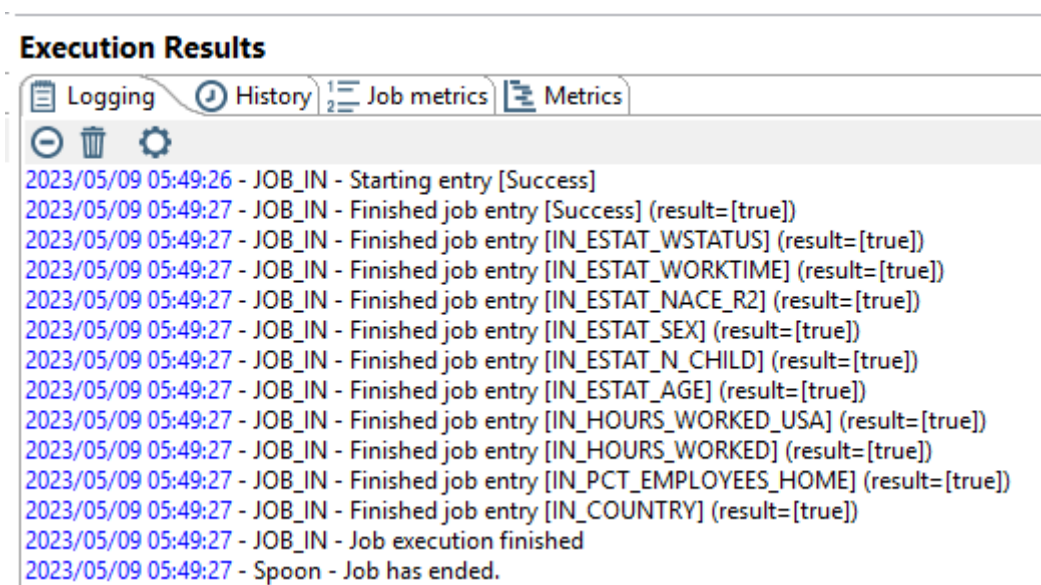
En aquest exemple, s'ha mostrat un cas bàsic de càrrega de dades, però, segons el format d'origen de les dades i de la seva qualitat, tal vegada sigui necessari utilitzar altres transformacions. Spoon disposa d'una gran quantitat de components, als quals es pot accedir des del menú lateral i que estan organitzats per categories.

3. Implementació dels treballs amb processos ETL

Els blocs de processos ETL implementats que cal tenir en compte són els següents:

- **Bloc IN:** processos ETL de transformació i càrrega a l'àrea intermèdia.
- **Bloc TR_DIM:** processos ETL de transformació i càrrega de dimensions.
- **Bloc TR_FACT:** processos ETL de transformació i càrrega de fets.
- **Procés complet:** utilitza els Jobs IN (Bloc IN) i TR (Bloc TR_DIM i TR_FACT)).

En aquest punt, per fer la càrrega efectiva de les dades, l'estudiantat ha de dissenyar mitjançant PDI els treballs que permetin l'execució seqüencial de tots els processos ETL inclosos en cada bloc. En aquest apartat s'han d'incloure també les volumetries obtingudes (nombre de registres carregats en cada taula) i **una captura de pantalla de la pestanya Logging** similar a la que es mostra a continuació:



Format i data de lliurament

El lliurament final d'aquesta activitat es farà a través de l'enllaç «Lliurament PR2» de l'espai *Continguts de l'aula*, adjuntant un únic fitxer amb la solució de la PR2, en format DOCX o PDF. El nom de l'arxiu a enviar estarà format per la composició del nom d'usuari i «_BDA_PR2». Per exemple, si el nom d'usuari és «bantich», el nom de l'arxiu ha de ser «bantich_BDA_PR2.pdf».

La data màxima de lliurament és el 25/12/2023 a les 23.59 hores.