

Guia i consells per al desenvolupament d'ETL.

Autor: José Luis Gómez

| | |
|--|----|
| Introducció. Guia i consells per al desenvolupament d'ETL | 2 |
| Esquema model conceptual: | 2 |
| Esquema model lògic: | 3 |
| Estructura de dades d'entrada: | 4 |
| 1. Matriu de Dimensions i Mètriques | 6 |
| Exemple Matriu Dimensions i Mètriques | 6 |
| 2. Simple i ordenat | 8 |
| Exemple patró de noms: | 8 |
| 3. Guia de desenvolupament d'ETL | 9 |
| Bloc <i>Staging</i> - Taules staging | 9 |
| Bloc Explotació- Taules DIM i FACT | 9 |
| 4. Plantilla Documentació Transformacions. | 11 |
| 5. Bones Practiques, configuració de l'entorn Spoon | 12 |
| Variables d'entorn | 12 |
| Connexió a la base de dades SQL Server | 13 |
| 6. EXEMPLE Transformació: IN_ENTRADA_UVA | 14 |
| Pas IN_UVA_Llegeix_Març | 14 |
| Pas IN_UVA_esborra_null | 16 |
| Pas IN_UVA_Normalització | 17 |
| Pas IN_UVA_Separa_Productor | 17 |
| Pas STG_UVA_CAMPANYA | 18 |
| Transformació completa | 19 |
| Consulta en la Base de dades. | 19 |
| Bibliografia | 20 |

Introducció. Guia i consells per al desenvolupament d'ETL

En aquesta guia es pretén recollir un conjunt de bones pràctiques relatives a la implementació d'un sistema de *data warehousing*, que, encara que no es refereixen als conceptes fonamentals de modelització, si poden resultar útils tant en la vida professional, com en la resolució i documentació de les Pràctiques i Treballs relatius al disseny, implementació i explotació de Bases de dades analítiques.

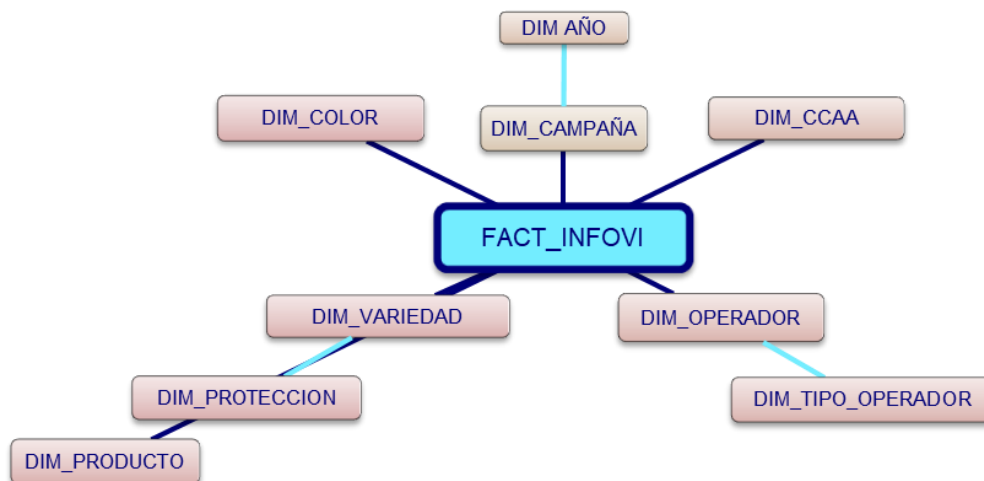
Tal com el seu nom indica, aquesta guia ha estat creada per a donar unes pautes sobre com abordar la part de creació dels processos d'extracció, transformació i càrrega (ETL) dels sistemes basats en un *data warehouse*.

Per a exemplificar els conceptes i bones pràctiques, s'utilitzarà dades de les campanyes vitivinícoles facilitades pel ministeri d'agricultura a través d'una eina específica per a la seva publicació i difusió, el Sistema d'Informació de Mercats del Sector Vitivinícola (INFOVI).

A continuació, per a situar-nos en el cas d'exemple que ens ocupa, es mostren els esquemes de model conceptual i lògic, així com el format de les fonts de dades d'entrada a integrar en el *data warehouse*.

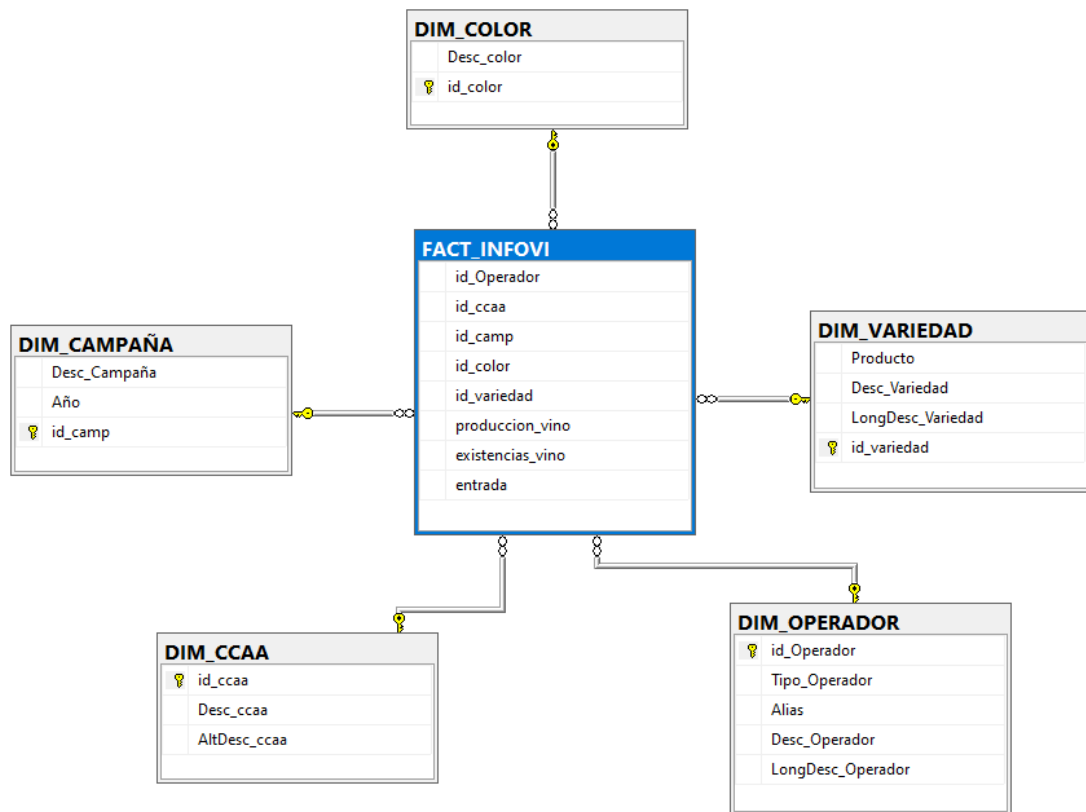
Esquema model conceptual:

Observem un model en estrella amb totes les seves dimensions:



Esquema model lògic:

La seva traducció a model lògic relacional seria:



Estructura de dades d'entrada:

L'estructura de les dades d'entrada prové majoritàriament de fulls de càlcul formatats. Cada FULL "entrada raïm" presenta un format com aquest:

Bloc TOTAL:

| | A | B | C | D |
|----|--|----------------------|----------------------|----------------------|
| 2 | CUADRO 1. ENTRADA DE UVA POR CCAA Y COLOR | | | |
| 3 | DE 1 DE AGOSTO DE 2019 A 30 DE NOVIEMBRE DE 2019 (kg) | | | |
| 4 | CCAA | Tinta | Blanca | Total |
| 5 | ANDALUCIA | 8.758.102 | 131.321.338 | 140.079.440 |
| 6 | ARAGON | 94.922.879 | 17.951.534 | 112.874.413 |
| 7 | ASTURIAS | 56.864 | 26.224 | 83.088 |
| 8 | BALEARES | 5.798.372 | 3.077.279 | 8.875.651 |
| 9 | CANARIAS | 2.703.374 | 4.106.610 | 6.809.984 |
| 10 | CANTABRIA | 93.449 | 54.278 | 147.727 |
| 11 | CASTILLA LA MANCHA | 1.118.386.092 | 1.533.589.873 | 2.651.975.965 |
| 12 | CASTILLA Y LEÓN | 147.877.642 | 121.043.656 | 268.921.298 |
| 13 | CATALUÑA | 102.713.678 | 328.459.108 | 431.172.786 |
| 14 | EXTREMADURA | 109.722.245 | 285.610.724 | 395.332.969 |
| 15 | GALICIA | 14.395.937 | 51.145.059 | 65.540.996 |
| 16 | C.MADRID | 5.259.063 | 2.960.399 | 8.219.462 |
| 17 | MURCIA | 96.684.586 | 5.114.589 | 101.799.175 |
| 18 | NAVARRA | 65.434.474 | 9.464.353 | 74.898.827 |
| 19 | PAIS VASCO | 72.190.027 | 14.782.552 | 86.972.579 |
| 20 | LA RIOJA | 274.611.447 | 31.529.945 | 306.141.392 |
| 21 | C.VALENCIANA | 241.610.283 | 97.837.042 | 339.447.325 |
| 22 | TOTAL | 2.361.218.514 | 2.638.074.563 | 4.999.293.077 |
| 23 | <i>Fuente: INFODIV, extracción de 08 de enero de 2020. Elaboración de SGF/HACIV a partir de datos de AICA, MAPA.</i> | | | |
| 24 | <i>NOTA Incluye la entrada de uva declarada por los productores de producción media de las últimas campañas ≥ 1.000 Hl y los de < 1.000 Hl. Es</i> | | | |
| 25 | <i>decir del conjunto de los productores</i> | | | |

Bloc Grans Productors:

| | | | | |
|----|--|----------------------|----------------------|----------------------|
| 25 | | | | |
| 26 | | | | |
| 27 | CUADRO 1.a ENTRADA DE UVA POR CCAA Y COLOR | | | |
| 28 | DE 1 DE AGOSTO DE 2019 A 30 DE NOVIEMBRE DE 2019 (kg). | | | |
| 29 | PRODUCTORES PROD. MEDIA ≥ 1000 HL | | | |
| 30 | CCAA | Tinta | Blanca | Total |
| 31 | ANDALUCIA | 6.146.521 | 129.803.938 | 135.950.459 |
| 32 | ARAGON | 92.796.026 | 17.648.506 | 110.444.532 |
| 33 | ASTURIAS | 0 | 0 | 0 |
| 34 | BALEARES | 3.403.728 | 2.214.643 | 5.618.371 |
| 35 | CANARIAS | 1.556.139 | 2.984.542 | 4.540.681 |
| 36 | CANTABRIA | 0 | 0 | 0 |
| 37 | CASTILLA LA MANCHA | 1.115.430.454 | 1.532.955.867 | 2.648.386.321 |
| 38 | CASTILLA Y LEÓN | 132.436.594 | 119.168.464 | 251.605.058 |
| 39 | CATALUÑA | 92.300.843 | 322.699.924 | 415.000.767 |
| 40 | EXTREMADURA | 108.968.168 | 284.756.446 | 393.724.614 |
| 41 | GALICIA | 9.190.331 | 42.708.796 | 51.899.127 |
| 42 | C.MADRID | 4.079.889 | 2.712.364 | 6.792.253 |
| 43 | MURCIA | 95.314.878 | 5.003.332 | 100.318.210 |
| 44 | NAVARRA | 64.238.117 | 9.332.804 | 73.570.921 |
| 45 | PAIS VASCO | 66.904.546 | 12.327.283 | 79.231.829 |
| 46 | LA RIOJA | 269.028.405 | 30.709.785 | 299.738.190 |
| 47 | C.VALENCIANA | 239.105.092 | 97.178.970 | 336.284.062 |
| 48 | TOTAL | 2.300.899.731 | 2.612.205.664 | 4.913.105.395 |
| 49 | <i>Fuente: INFODIV, extracción de 08 de enero de 2020. Elaboración de SGF/HACIV a partir de datos de AICA, MAPA.</i> | | | |

Bloc Petits Productors:

| | | | | |
|----|---|-------------------|-------------------|-------------------|
| 49 | | | | |
| 50 | | | | |
| 51 | CUADRO 1.b ENTRADA DE UVA POR CCAA Y COLOR DE 1 DE AGOSTO DE 2019 A 30 DE NOVIEMBRE DE 2019 (kg). PRODUCTORES PROD. MEDIA < 1000 HL | | | |
| 52 | | | | |
| 53 | CCAA | Tinta | Blanca | Total |
| 54 | ANDALUCIA | 2.611.581 | 1.517.400 | 4.128.981 |
| 55 | ARAGON | 2.126.853 | 303.028 | 2.429.881 |
| 56 | ASTURIAS | 56.864 | 26.224 | 83.088 |
| 57 | BALEARES | 2.394.644 | 862.636 | 3.257.280 |
| 58 | CANARIAS | 1.147.235 | 1.122.068 | 2.269.303 |
| 59 | CANTABRIA | 93.449 | 54.278 | 147.727 |
| 60 | CASTILLA LA MANCHA | 2.955.638 | 634.006 | 3.589.644 |
| 61 | CASTILLA Y LEÓN | 15.441.048 | 1.875.192 | 17.316.240 |
| 62 | CATALUÑA | 10.412.835 | 5.759.184 | 16.172.019 |
| 63 | EXTREMADURA | 754.077 | 854.278 | 1.608.355 |
| 64 | GALICIA | 5.205.606 | 8.436.263 | 13.641.869 |
| 65 | C.MADRID | 1.179.174 | 248.035 | 1.427.209 |
| 66 | MURCIA | 1.369.708 | 111.257 | 1.480.965 |
| 67 | NAVARRA | 1.196.357 | 131.549 | 1.327.906 |
| 68 | PAIS VASCO | 5.285.481 | 2.455.269 | 7.740.750 |
| 69 | LA RIOJA | 5.583.042 | 820.160 | 6.403.202 |
| 70 | C.VALENCIANA | 2.505.191 | 658.072 | 3.163.263 |
| 71 | TOTAL | 60.318.783 | 25.868.899 | 86.187.682 |
| 72 | Fuente: INFODIV, extracción de 08 de enero de 2020. Elaboración de SGRHACIV a partir de datos de AICA, MAPA. | | | |
| 73 | | | | |

A partir de la informació descrita fins aquí, es donaran algunes recomanacions per a la creació de processos ETL aplicats a aquest cas, que us poden ser útils cara a la realització de la PEC2 i la PRA2 del curs.

1. Matriu de Dimensions i Mètriques

La Matriu de Dimensions i Mètriques (Kimball, *Data Warehouse Toolkit*, 2013), és una eina clau en el disseny del *data warehouse*, que representa els processos centrals de l'organització i la dimensionalitat associada. Aquest model busca proporcionar la perspectiva necessària per a garantir, que tota l'empresa, pugui integrar les seves dades a l'entorn del *data warehouse*.

Una dels seves principals avantatges és que permet combinar característiques del model conceptual i lògic, amb característiques a tenir en compte en la implementació de l'ETL.

Partint de l'anàlisi de requeriments i disseny del *data warehouse* (PRA1), aquesta eina ha de servir d'ajuda per a la creació de processos ETL que permetin la càrrega del *data warehouse* (PRA2). La seva finalitat és ajudar a organitzar les idees de manera que es pugui tenir una visió completa sobre com es relacionen aquests processos i quines transformacions són necessàries per a arribar al model dimensional buscat.

S'ha de crear una matriu, en la que es representin les Dimensions en files i els fets en columnes. En intersecció fila-columna, s'anotarà la informació rellevant referent a aquesta relació. Vegem com es procediria amb el cas concret InfoVi.

Exemple Matriu Dimensions i Mètriques

Segons l'estructura de les dades d'entrada, el full de càlcul "entrada raïm" que recull les dades d'entrada, està estructurada en tres blocs d'informació. El primer bloc conté informació de resum, redundant, que es pot descartar. Els altres dos, es refereixen als operadors, distingint entre grans i petits productors.

Operador: El valor de l'Operador apareix en la capçalera de cada bloc com a part d'un text major. Com que la seva localització i extracció és difícil, es recomana prendre el valor segons la posició que ocupa el bloc. És a dir, a partir de la línia 29 per a grans i 53 per a petits.

Dates de campanya: Aquest valor també apareix de forma no completament estructurada, com a part de l'encapçalat del cada bloc. Una opció més simple és prendre-la del nom del fitxer, ja que té el següent format:

INFOVI_YYMM.xlsx on YYYY=any numèric 4 dígits i MM=mes numèric 2 dígits, per exemple INFOVI_201911.xlsx

Color: Els valors d'aquesta dimensió (Tinta, Blanca), apareixen en columnes, per la qual cosa serà necessari normalitzar-les¹ per a la seva integració en el model dimensional

¹ Equivalent a *unpivot* en SQL o Excel, convertir columnes a files

La Matriu de Dimensions i Mètriques per al cas que ens ocupa, podria ser la següent:

| PROCÉS | ENTRADA_UVA | PROD_VI_DESG | EXIST_VI_DESG |
|----------------------------|---|--|--|
| DIM\Mètrica | Entrada Raïm (Kg) | Producció (Hl). | Existències Vi (Hl). |
| Campanya | Extreure a partir de nom fitxer | Extreure a partir de nom fitxer | Extreure a partir de nom fitxer |
| CCAA | X | X | X |
| Color | Apareix en columnes, normalitzar | Apareix en columnes, normalitzar | Apareix en columnes, normalitzar |
| Tipus Operador | A partir de posició: Total, Grans, Petits | A partir de la posició: Total, Grans, Petits | A partir de la posició: Total, Grans, Petits, Operador de magatzem |
| Categoria Protecció | Varietat | N/A | X |

En aquesta matriu, les columnes representen processos importants del sector vitivinícola, concretament el procés d'entrada de raïm, focus d'aquest exemple, i els processos de producció i existències de vi.

Cada procés té associada una les seves pròpies mètriques: Entrada, producció i existències.

En les files es representen les dimensions identificades en el model conceptual.

La intersecció de les files i columnes (Dimensió i Procés-Mètrica), indica si estan relacionades, i conté informació útil tant relativa a la relació en si, com a consideracions en l'ETL:

- Campanya: en tots els processos, el valor s'extraurà del nom del fitxer.
- CCAA: està present en tots els processos de manera directa.
- Color: apareix com encapçalat de columna, en tots els casos.
- Operador: Només en existències apareix l'operador de magatzem.
- Varietat: No existeix (N/A) en Entrada Raïm.

2. Simple i ordenat

Com sabem la majoria de sistemes de data *warehousing* neixen de necessitats de negoci, que requereixen prendre decisions a partir d'una anàlisi prèvia i profunda de les dades. Atès que els negocis evolucionen, també evolucionen les necessitats d'informació, i els nivells d'exigència cap a aquestes.

Aquesta situació provoca no només l'actualització continua del desenvolupament d'aquests sistemes, sinó també, el fet que diferents persones, en diferents moments col·laborin en aquests projectes, per la qual cosa és absolutament necessari establir una sèrie de regles i normes, amb l'objectiu de minimitzar el temps necessari per a familiaritzar-se amb un desenvolupament.

A més a més, en moltes ocasions, el nombre d'elements dels sistemes (processos ETL, taules, atributs, mètriques, funcions, algorismes, documentació entre altres) creixen fins a fer-se complexos i ingovernables; sent difícil de localitzar i entendre un objecte concret, fins i tot dins d'un mateix projecte.

Encara que existeixen eines específiques per a aquesta gestió, una bona organització dels elements, continguts i accions, ajudarà a localitzar i contextualitzar els objectes utilitzats, permetent que el sistema es mantingui simple i ordenat.

Per a dur-ho a terme, durant tot el disseny i implementació cal utilitzar una metodologia que permeti definir tant les regles per a nomenar de manera simple i organitzada els elements del sistema (**patró de noms**) com l'estructura que ha de seguir la implementació (**guia**) o fins i tot la documentació (**plantilla**).

Mitjançant aquestes regles s'ha de buscar:

- Reduir: La simplicitat consisteix a restar l'obvi i agregar el més significatiu, la forma més senzilla d'aconseguir la simplicitat és mitjançant una reducció acurada.
- Retolar, nomenar, posar etiquetes. Utilitzar elements coneguts i fàcilment interpretables de manera correcta per l'usuari, gràcies a models mentals ja creats.
- Contextualitzar: Utilitzar elements coneguts i fàcilment interpretables de manera correcta.
- Integrar: agrupar aquells elements relacionats per a reduir el nombre de categories.
- Prioritzar o jerarquitzar en els diferents nivells de la nostra taxonomia, donant més importància als elements principals.

Exemple patró de noms:

Taules STAGING

STG.IN_<extensió>_Nom

STG.IN_json_CCAA

Taules DIMENSIÓ

Prod.DIM_Nom_menor_jerarquia

PROD.DIM_Dia

COLUMNES de taula Dimensió

Nom_jerarquia_<ID | DESC> identificació (ID) i descripció (DESC)

Dia_ID

3. Guia de desenvolupament d'ETL.

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 tres etapes: extracció, transformació i càrrega.

Bloc *Staging*- Taules *staging*

L'*STAGING AREA* és una àrea intermèdia d'emmagatzematge de dades utilitzada per al processament d'aquestes durant els processos d'extracció, transformació i càrrega (ETL). Per a extreure informació de diferents fonts i carregar-la en aquesta àrea, s'utilitzen taules intermèdies o *STAGING*. Es recomana començar carregant tots els orígens de dades per tipus, això és generant una taula *staging* específica per a tots els orígens de dades d'un mateix tipus.

L'esquema de passos necessaris per a implementar cada JOB per origen de dades:

- 1) Lectura del fitxer.
- 2) Transformacions:
 - a. Netejar valors nuls.
 - b. Eliminar Duplicats.
 - c. Normalitzar columnes.
 - d. Ordenar
 - e. ...
- 3) Càrrega a la taula intermèdia.

Bloc Explotació- Taules DIM i FACT

Un altre tipus de taules que s'han de carregar amb dades processades finals són aquelles que seran explotades en el model dimensional. Segons sigui la seva funció en el model dimensional, existeixen dos tipus: taules de dimensions (DIM) i taules de fets (FACT). Des del punt de vista del procés de l'ETL, tenen una sèrie de passos comuns i altres específics que es detallen en els següents blocs.

Bloc Passos comuns:

- 1) Lectura de la taula IN_XXX, on es seleccionen els camps a utilitzar.
- 2) Transformacions:
 - a. Assegurar valors únics,
 - b. Canviar de nom columnes
 - c. ...

Bloc Taules de dimensions (DIM):

- 1) Cerca de nous valors de dimensió: amb el propòsit d'identificar nous valors de dimensió, es comparen les dades de dimensió de les taules de Staging, amb les dades ja existents (de producció en el cas d'exemple). Cal tenir en compte que no tenir un valor (*null*) també és un valor vàlid, per la qual cosa pot ser necessari crear un valor per defecte del tipus: "N/D", "01/01/9999",.....
- 2) Creació de les claus primàries: per a donar d'alta els nous valors en una taula de dimensió és necessari generar un nou id únic (mitjançant autonumèric), en producció. És important conservar addicionalment el valor de l'atribut com ve en *Staging* (valor *Staging*), ja que després serà necessari en carregar la FACT i cercar els ID's que referència.
- 3) Càrrega a producció (explotació) de la dimensió: Es guarda tot el procés en una taula de dimensió DIM_XXXX.

Bloc Explotació Fets (FACT):

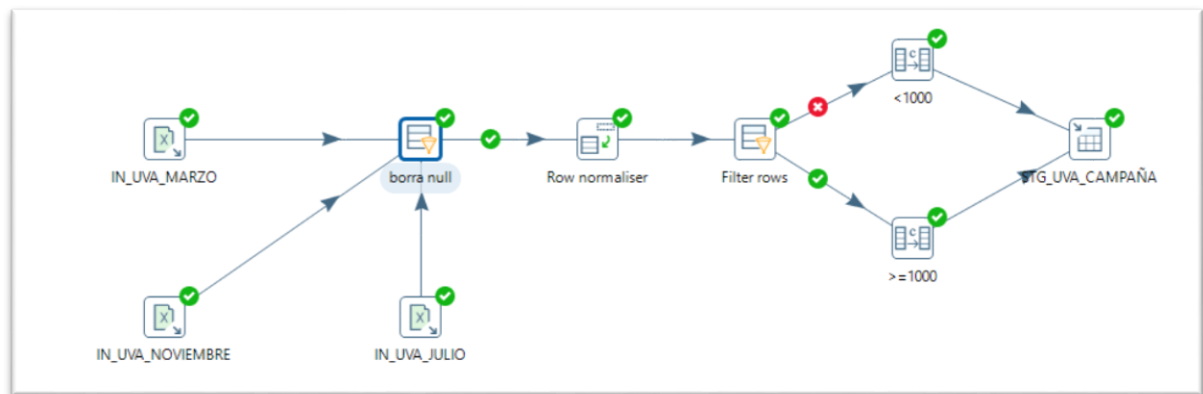
- 1) Transformació del "valor Staging" a ID de la dimensió: una vegada identificades les mètriques i fets de les taules de l'*Staging*, aquestes han de relacionar-se amb les taules DIM_ de producció, per a això s'utilitza com a nexa d'unió el "valor Staging" de la dimensió. En la taula de fet només és necessari que aparegui l'ID de la dimensió a la qual referència, sense tenir necessitat d'incloure altres descripcions o el "valor Staging".
- 2) Càrrega a producció (explotació) del fet: localitzades les "FACT" de l'*Staging* i els ID de Dimensió, es procedeix a la inserció en les Taules FACT de producció (FACT_XXXX).

4. Plantilla Documentació Transformacions.

Spoon és el dissenyador gràfic de transformacions i treballs del sistema d'ETL.

Els treballs (Job) controlen el flux a alt nivell, això és, concatenen de manera seqüencial i ordenada les transformacions, enviament de correus en cas de fallida, transferir arxius a través de FTP,....

Les transformacions consisteixen a moure i transformar files des de l'origen a la destinació, per a això es concatenen tant seqüencial com paral·lelament una sèrie de passos amb una lògica o programació parcial que en conjunt realitzen la transformació completa.



A l'hora de documentar una transformació ETL, i especialment per a la seva documentació en la PRAC2, el document ha d'incloure, informació de cadascun dels passos involucrats, així com de la seva seqüència en la transformació:

- La descripció de cada PAS ha de tenir l'esquema següent:
 1. Nom del component *Spoon* i del pas específic.
 2. Descripció del component i del pas.
 3. Paràmetres: pestanyes i paràmetres necessaris.
 4. Informació importada, com els camps importats mitjançant la pestanya "Fields".
 5. Preview de les dades obtingudes en el pas.
- Diagrama del job executat, incloent les dades de la pestanya "step metrics".
- Consulta en la Base de dades on es mostrin les primeres línies.

5. Bones Pràctiques, configuració de l'entorn Spoon

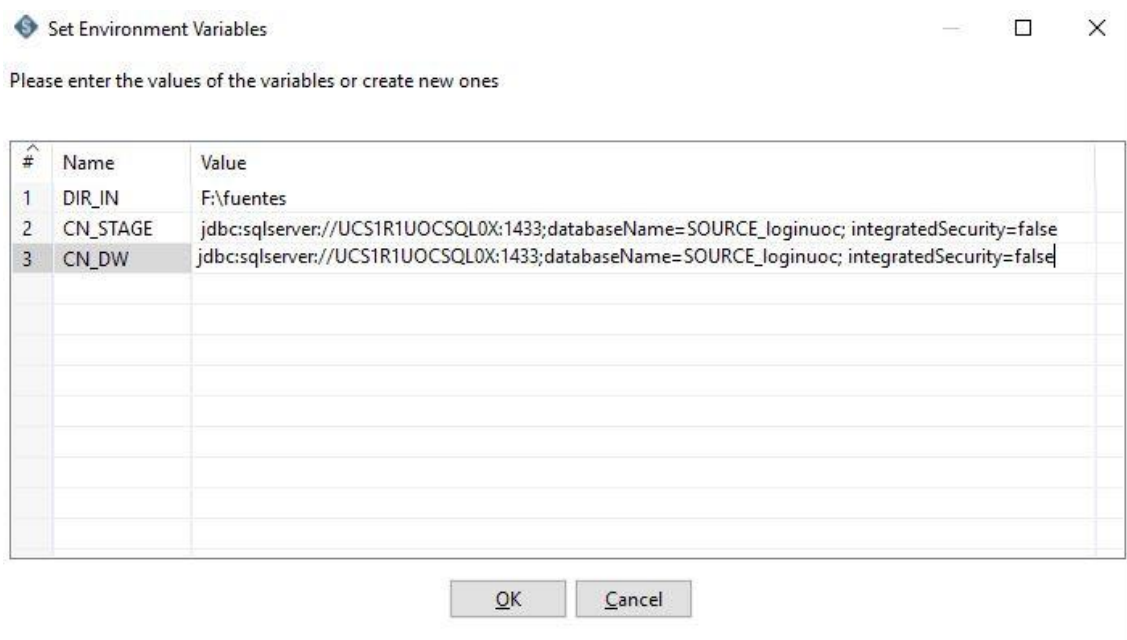
Variables d'entorn

És una bona pràctica utilitzar variables d'entorn per a evitar introduir errors en definicions repetitives durant la implementació dels processos. PDI us permet afegir variables personalitzades i pròpies dels vostres desenvolupaments en l'arxiu «kettle.properties».

Per a les PRAC, són necessàries tres variables. Una per a emmagatzemar la ruta de les fonts de dades i altres dues per a configurar les cadenes de connexió a la base de dades, «CN_STAGE» (àrea intermèdia / *staging area*) i «CN_DW» (*data warehouse*).

| Variable | Valor |
|----------|---|
| DIR_IN | F:\fuentes |
| CN_STAGE | jdbc:sqlserver://UCS1R1UOCSQL0X:1433;databaseName=SOURCE_loginuoc; integratedSecurity=false |
| CN_DW | jdbc:sqlserver://UCS1R1UOCSQL0X:1433;databaseName=SOURCE_loginuoc; integratedSecurity=false |

La referència a les variables d'entorn durant la implementació dels processos es realitza mitjançant claus, d'aquesta manera: {DIR_IN}, {CN_STAGE}, {CN_DW}.



Set Environment Variables

Please enter the values of the variables or create new ones

| # | Name | Value |
|---|----------|---|
| 1 | DIR_IN | F:\fuentes |
| 2 | CN_STAGE | jdbc:sqlserver://UCS1R1UOCSQL0X:1433;databaseName=SOURCE_loginuoc; integratedSecurity=false |
| 3 | CN_DW | jdbc:sqlserver://UCS1R1UOCSQL0X:1433;databaseName=SOURCE_loginuoc; integratedSecurity=false |

OK Cancel

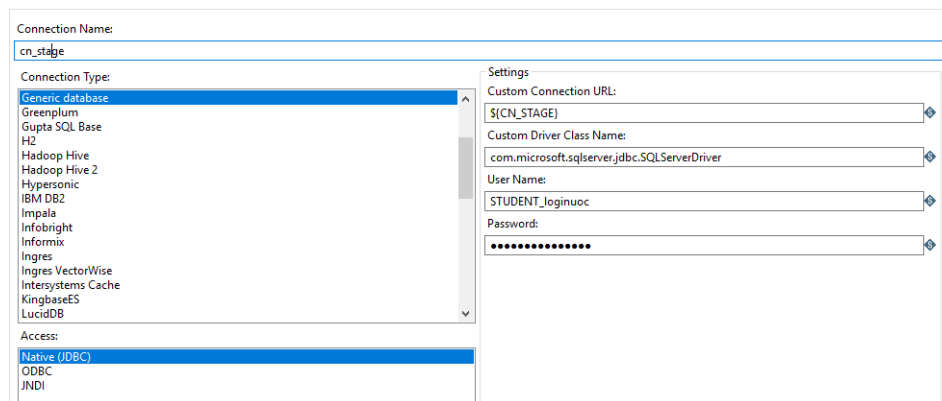
Connexió a la base de dades SQL Server

Un altre pas previ que s'ha de realitzar és la creació de les connexions a les bases de dades que s'usen en totes les transformacions i treballs dels processos de càrrega.

S'han definit dues connexions diferents, una per a la base de dades del model multidimensional («BBDD») i una altra per a l'àrea intermèdia («STAGE»); d'aquesta manera es diferencia clarament el seu ús, encara que físicament es refereixin al mateix esquema de la base de dades.

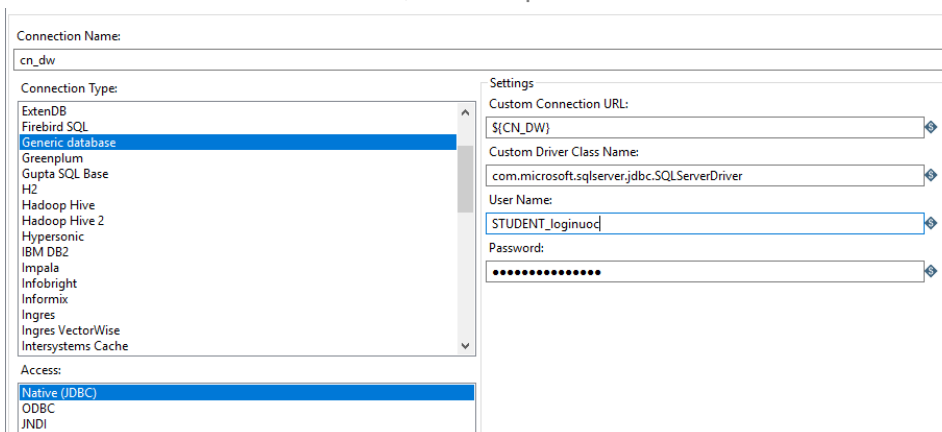
Es podria crear un esquema *stage* en l'SQL Server dins de la base de dades assignada a l'estudiant per a carregar les taules intermèdies (IN_) i definir la variable «CN_STAGE» fent referència a aquest esquema, però per a simplificar la solució de la pràctica es carregaran totes les taules a l'esquema per defecte, dbo.

En la creació de la connexió al «STAGE», el nom que utilitzarem és «cn_stage»:



The screenshot shows a configuration window for a new connection named 'cn_stage'. The 'Connection Type' is set to 'Generic database'. The 'Access' dropdown is set to 'Native (JDBC)'. The 'Settings' section on the right includes: 'Custom Connection URL' set to '{CN_STAGE}', 'Custom Driver Class Name' set to 'com.microsoft.sqlserver.jdbc.SQLServerDriver', 'User Name' set to 'STUDENT_loginuoc', and 'Password' masked with dots.

En la creació de la connexió al «DW», el nom que li donarem és «cn_dw»:



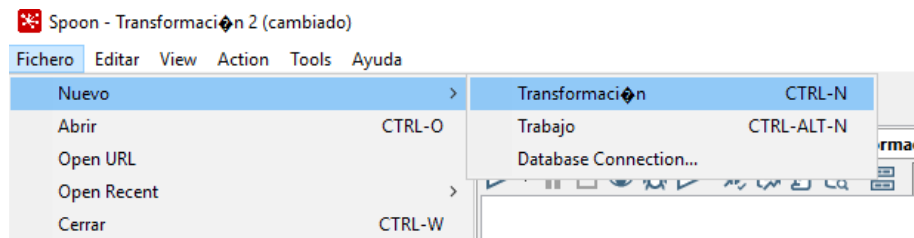
The screenshot shows a configuration window for a new connection named 'cn_dw'. The 'Connection Type' is set to 'Generic database'. The 'Access' dropdown is set to 'Native (JDBC)'. The 'Settings' section on the right includes: 'Custom Connection URL' set to '{CN_DW}', 'Custom Driver Class Name' set to 'com.microsoft.sqlserver.jdbc.SQLServerDriver', 'User Name' set to 'STUDENT_loginuoc', and 'Password' masked with dots.

6. EXEMPLE Transformació: IN_ENTRADA_UVA

El següent apartat pretén mostrar l'aplicació pràctica de la combinació de:

- Guia desenvolupament ETL, bloc STAGING
- Plantilla de Documentació Transformacions

No tots els apartats estan totalment coberts i només s'inclouen els passos més importants. En el cas de les PRAC de l'assignatura serà necessària una explicació introductòria per contextualitzar el desenvolupament.



Pas IN_UVA_Llegeix_Març

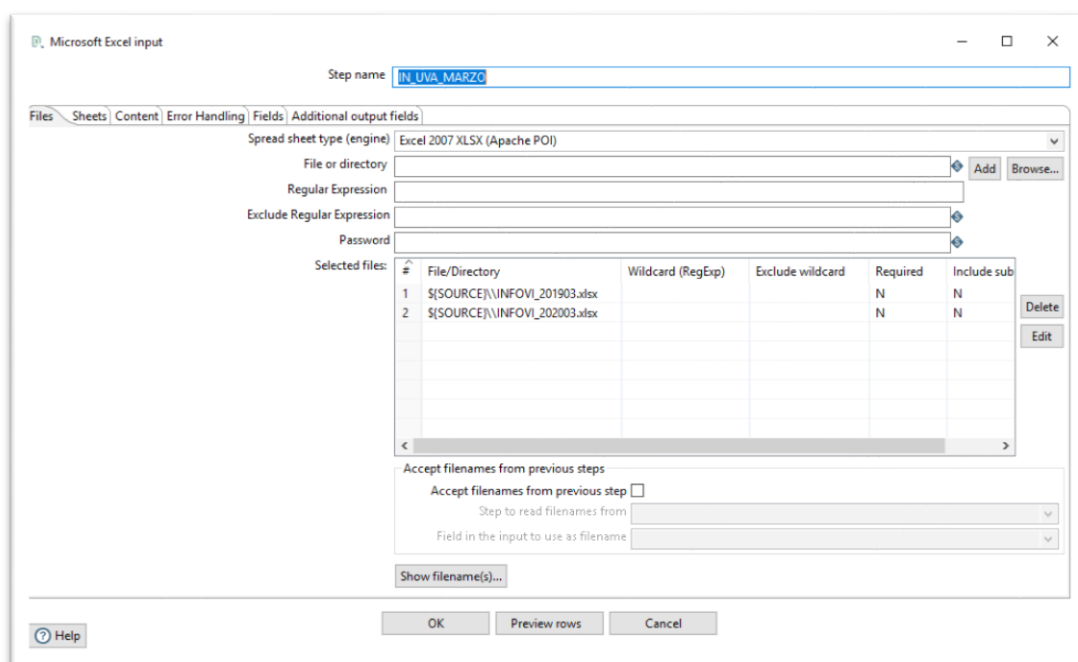
Component: Microsoft Excel input

Descripció: Permet carregar les dades d'entrada provinents d'un fitxer Excel. Carrega l'entrada de raïm dels dos Excel de les campanyes de març.

El primer pas de la transformació correspon a la lectura del fitxer origen, com es tracta d'un fitxer XLSX s'utilitzarà com a entrada el tipus «Microsoft Excel Input».

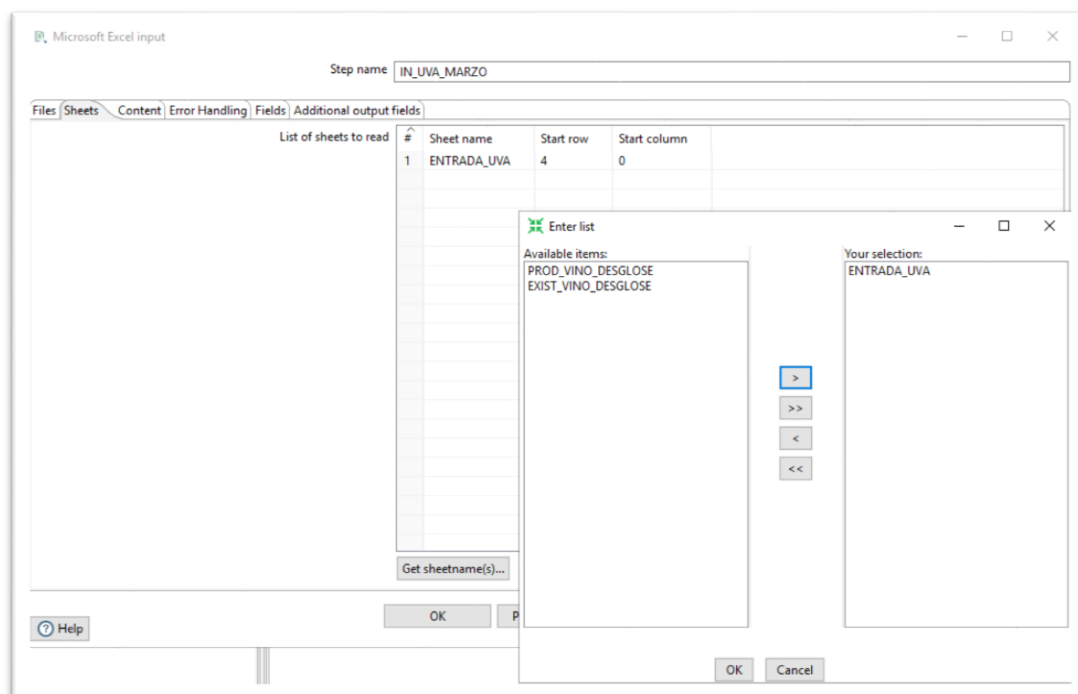
Paràmetres:

FILES: "file or directory": Per a facilitar la lectura del fitxer s'utilitza la variable d'entorn «SOURCES».



Sheets; mitjançant “get sheet names”, full “ENTRADA_UVA”, amb “start row” = 4.

Per a la resta de les campanyes, la fila inicial no haurà d’incloure la taula d’agregats inicials.



Fields Mitjançant el botó «Get fields from header row...» s’obtenen tots els camps del fitxer, així com el tipus, format i longitud de la dada. És necessari revisar que siguin correctes. La línia 4 fa referència a la columna de Totals i no aporta dades noves

Microsoft Excel input

Step name:

| # | Name | Type | Length | Precision | Trim type | Repeat | Format | Currency | Decimal | Grouping |
|---|--------|--------|--------|-----------|-----------|--------|--------|----------|---------|----------|
| 1 | CCAA | String | -1 | -1 | none | N | | | | |
| 2 | Tinta | Number | -1 | -1 | none | N | | | | |
| 3 | Blanca | Number | -1 | -1 | none | N | | | | |
| 4 | Total | Number | -1 | -1 | none | N | | | | |

Get fields from header row...

Help OK Preview rows Cancel

Preview: botó «Preview rows» .

Examine preview data

rows of step: Microsoft Excel input (40 rows)

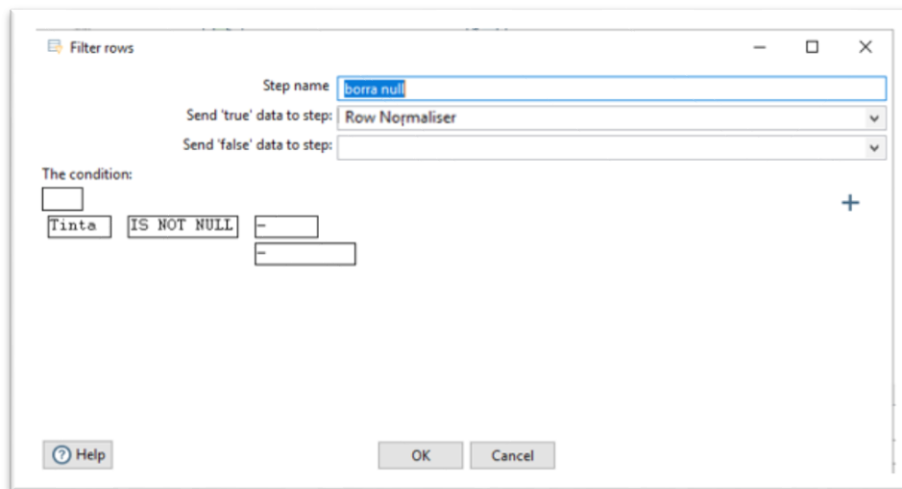
| # | CCAA | Tinta | Blanca |
|----|---|--------------|--------------|
| 1 | ANDALUCIA | 7382625.0 | 163111455.0 |
| 2 | ARAGON | 161676180.0 | 23234898.0 |
| 3 | ASTURIAS | 0.0 | 0.0 |
| 4 | BALEARES | 3735996.0 | 2380115.0 |
| 5 | CANARIAS | 2075169.0 | 4356001.0 |
| 6 | CANTABRIA | 0.0 | 0.0 |
| 7 | CASTILLA LA MANCHA | 1397381799.0 | 2342200051.0 |
| 8 | CASTILLA Y LEÓN | 173551710.0 | 138289129.0 |
| 9 | CATALUÑA | 95035408.0 | 327703746.0 |
| 10 | EXTREMADURA | 136650394.0 | 359490274.0 |
| 11 | GALICIA | 8008966.0 | 44187919.0 |
| 12 | C.MADRID | 8761583.0 | 6095307.0 |
| 13 | MURCIA | 113981527.0 | 5604018.0 |
| 14 | NAVARRA | 95623746.0 | 13199850.0 |
| 15 | PAIS VASCO | 89981864.0 | 14507533.0 |
| 16 | LA RIOJA | 294267938.0 | 43904079.0 |
| 17 | C.VALENCIANA | 236540058.0 | 94673040.0 |
| 18 | TOTAL | 2824654963.0 | 3581936315.0 |
| 19 | Fuente: INFOVI, extracción de 6 de mayo de 2019. Elaboración de SGFHV a partir de datos de AICA, MAPA. | | |
| 20 | NOTA Incluye la entrada de uva declarada por los productores de producción media de las últimas campañas ≥ 1.000 HI | | |
| 21 | ANDALUCIA | 6146521.0 | 129803938.0 |
| 22 | ARAGON | 92869176.0 | 17679456.0 |
| 23 | ASTURIAS | 0.0 | 0.0 |
| 24 | BALEARES | 3403728.0 | 2214643.0 |
| 25 | CANARIAS | 1556139.0 | 2984542.0 |
| 26 | CANTABRIA | 0.0 | 0.0 |
| 27 | CASTILLA LA MANCHA | 1117488004.0 | 1538409157.0 |
| 28 | CASTILLA Y LEÓN | 135386623.0 | 121134786.0 |
| 29 | CATALUÑA | 92060774.0 | 322562168.0 |
| 30 | EXTREMADURA | 108968168.0 | 284756446.0 |
| 31 | GALICIA | 9243825.0 | 42885687.0 |
| 32 | C.MADRID | 4191439.0 | 2712364.0 |
| 33 | MURCIA | 95314878.0 | 5003332.0 |

Pas IN_UVA_esborra_null

Component: Filter Rows

Descripció: Filtra determinats Valors. Per exemple, per a descartar les files amb nuls, l'opció "filter rows", permet carregar o excloure dades en funció d'una condició.

Paràmetres:



Filter rows

Step name:

Send 'true' data to step:

Send 'false' data to step:

The condition:

IS NOT NULL

Send true data to step: Envia les dades que compleixen la condició al següent pas «Row Normaliser».

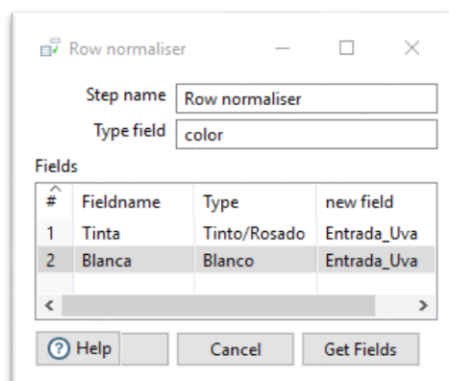
The Condition: Avalua que la columna «Tinta» no sigui nul.

Pas IN_UVA_Normalització

Component: Row normaliser

Descripció: Normalitza columnes de les nostres dades a una sola o vàries. En aquest cas, ens interessa que les columnes Tinta i Blanca ens les posi com a varietat. Aquest pas ens permetrà etiquetar-les com a entrada (quantificarà aquests camps).

Paràmetres:



Row normaliser

Step name:

Type field:

Fields

| # | Fieldname | Type | new field |
|---|-----------|--------------|-------------|
| 1 | Tinta | Tinto/Rosado | Entrada_Uva |
| 2 | Blanca | Blanco | Entrada_Uva |

Pas IN_UVA_Separa_Productor

Component: Filter Rows

Descripció: Filtrar valors. Segons la posició en el full Excel, discrimina el tipus de productor.

...

Pas STG_UVA_CAMPANYA

Component: Table Output

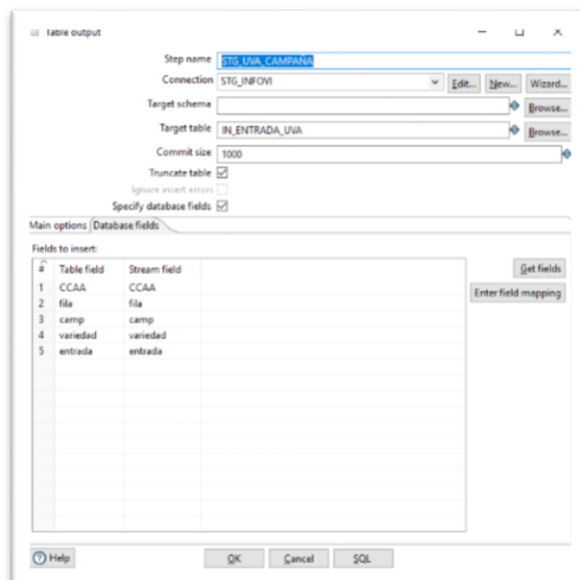
Descripció: carrega totes les dades dels passos anteriors a una taula.

Paràmetres:

Connection: STG_INFOVI

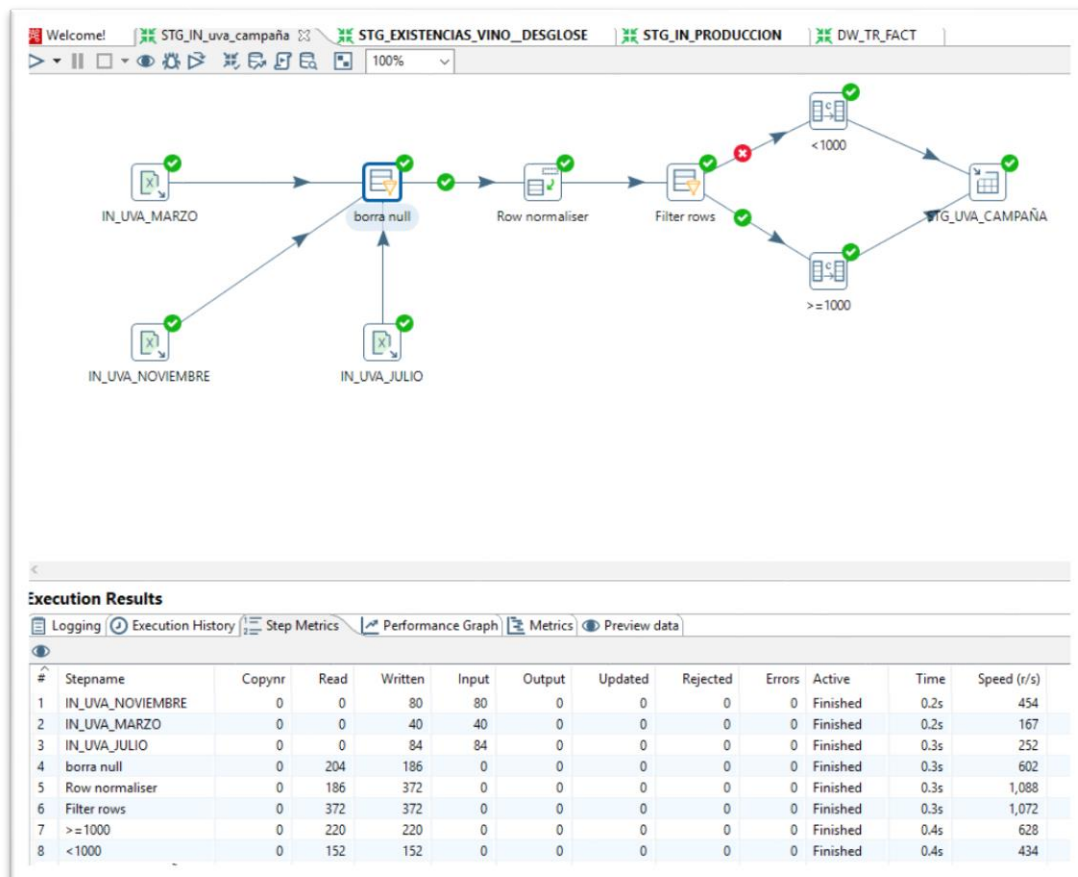
Target table: IN_ENTRADA_UVA

Truncate table: Facilita per a possibles reprocessaments



Transformació completa

La transformació completa és la següent, incloent la pestanya informativa d'execució "step metrics"



Consulta en la Base de dades.

Select * from ...

Bibliografia

Kimball and Margy Ross, The Data Warehouse Toolkit, 2013
<https://www.kimballgroup.com/data-warehouse-business-intelligence-resources/kimball-techniques/kimball-data-warehouse-bus-architecture/>

John Maeda, Laws of Simplicity, 2005. <http://lawsofsimplicity.com/>

Abby Covert, How to make sense of any mes, 2015
<http://www.howtomakesenseofanymess.com/>

Don Norman, Design as communication, 2004.
https://jnd.org/design_as_communication/