MAGATZEM DE DADES: Mobilitat d'estudiants Erasmus

Carlos Cabello Martin

Treball Final de Carrera
Enginyeria Tècnica en Informàtica de Gestió
UOC

Consultor: Bartomeu Antich Luque

01/2016


Aquesta obra està subjecta a una llicència de Reconeixement-NoComercial-SenseObraDerivada 3.0 Espanya de Creative Commons

FITXA DEL TREBALL FINAL

Títol del treball:	*Mobilitat d'estudiants Erasmus*
Nom de l'autor:	*Carlos Cabello Martin*
Nom del consultor:	*Bartomeu Antich Luque*
Data de lliurament (mm/aaaa):	*01/2016*
Àrea del Treball Final:	*Magatzem de dades*
Titulació:	*Enginyeria Tècnica en Informàtica de Gestió*

Resum del Treball (màxim 250 paraules):

Aquest projecte fa referència a l'àrea de magatzems de dades, que avui dia és una solució molt utilitzada al món real, tant a empreses privades com publiques. Estem parlant d'una base de dades amb informació històrica especialment dissenyada per a realitzar-hi consultes eficientment, per tal de prendre decisions importants a temps.

En concret en aquest projecte es vol una solució *open source* de BI que permeti fer una anàlisi acurada de la mobilitat d'estudiants d'Erasmus, incidint en la transferència d'alumnes entre països, Institucions que més estudiants envien i les que més reben...

Per dur a terme la solució s'ha creat un magatzem de dades que cobrís tots els conceptes dels fitxers d'entrada de dades. S'han transformat les dades mitjançant processos ETL i posteriorment s'ha creat un model multidimensional per a l'explotació les dades, utilitzant una plataforma d'informes web. El *software* utilitzat per fer totes aquestes tasques ha estat *BI Pentaho*.

La consecució del projecte ha cobert totes les expectatives i requeriments inicials. La solució està preparada per rebre dades de cursos vinents que es realitzin a més d'acceptar possibles modificacions en el model i informes sense un cost elevat.

Abstract (in English, 250 words or less):

This project is related to the data warehouse area which is widely used in the business world nowadays, both in private and public companies. We are talking about a historical database with information specifically designed to process queries efficiently in order to make important decisions on time.

To be precise we need a BI open source solution in this project which permits an accurate analysis about Erasmus student's mobility, focusing on the student's movements between countries belonging to the program, Universities who send the highest number of students and Universities at which the students spend their Erasmus period…

To achieve this solution, we have created a data warehouse in order to cover all the existing concepts contained in the source files. We have transformed the data using ETL processes. Also, we have created a multidimensional model in order to use the data in depth by means of a web-based platform. The selected software for doing these tasks has been *BI Pentaho*.

This Erasmus project has covered all expectations and initial requirements. The solution is already prepared to receive new data from future school years. Also, the application is capable of accepting new modifications in the model and reports at low cost.

Paraules clau (entre 4 i 8):

Magatzem de dades, *Pentaho*, *star-schema*, OLAP, ETL.

ÍNDEX

LLISTA DE FIGURES

LLISTA DE TAULES

1. INTRODUCCIÓ

1.1 Context i justificació del Treball

El treball final de carrera (TFC) que es presenta en aquest document consisteix a crear un magatzem de dades sobre la mobilitat dels estudiants d'Erasmus.

El programa Erasmus, que va ser instaurat l'any 1987, rep el seu nom de l'Anglès ***EuR**opean Community **A**ction **S**cheme for the **M**obility of **U**niversity **S**tudents.* És un pla d'acció de la Unió Europea per a la mobilitat d'estudiants, i el que persegueix és exactament el que diu el seu nom, la mobilitat acadèmica d'estudiants entre països.

Cada any augmenta el nombre d'estudiants que hi participen, segons dades de la Comissió Europea gairebé 270.000 estudiants van participar en el curs 2012-2013, fent que cada vegada hi hagi més informació a tractar. Aquest creixement fa necessària una solució que pugui gestionar totes aquestes dades.

Com a punt de partida tenim una sèrie de fitxers, fonts de dades, en format excel i csv. Hi ha una necessitat d'analitzar la informació, tasca molt complexa amb els actuals fitxers, però no impossible. Si es volgués, es podria fer una mena d'anàlisi amb taules dinàmiques.

La millor solució quan es treballa amb grans volums de dades és crear un magatzem de dades, crear un procés de càrrega ETL i explotar la informació des dels informes OLAP, "atacant" un model multidimensional. I això és el que s'ha fet en aquest projecte final de carrera.

1.2 Objectius del Treball

1.2.1 Generals

L'objectiu principal del projecte és adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible en una base de dades transaccional, així com la utilització d'eines que facilitin la generació d'informes.

El projecte ha de ser construït en un model multidimensional del tipus *Star-schema* i es tracta d'aprofundir en el disseny d'aquest model.

Dins d'aquest marc es tracta de desenvolupar els següents objectius:

- Analitzar un problema complex de tipus pràctic transformant-lo en un projecte informàtic.
- Planificar i estructurar el desenvolupament del projecte mitjançant l'elaboració d'un pla de treball aplicant una metodologia adient.
- Treballar a fons els aspectes formals del desenvolupament de projectes.
- Sintetitzar una solució viable i realista al problema proposat, basada en l'anàlisi de requeriments.
- Elaborar una memòria del projecte segons una estructura prefixada.
- Elaborar una presentació del desenvolupament i resultats finals del projecte.

1.2.2 Específics

L'objectiu específic d'aquest projecte és integrar totes les fonts de dades, que ara són fitxers csv i Microsoft Excel d'Erasmus. És a dir ser capaços, partint d'aquestes fonts de dades, de carregar-les, transformar-les, adaptar-les i oferir-les als analistes.

Per fer això es requereix usar diferents tecnologies de bases de dades i de modelatge que més endavant veurem.

1.3 Enfocament i mètode seguit

Com a tot projecte informàtic, és necessari seguir una sèrie de fases abans d'arribar a la solució final. Hi ha moltes maneres de fer un projecte, però en el nostre cas hem escollit el mètode de cicle de vida, que consta de les següents etapes:

- Estudi preliminar
- Determinació dels requeriments
- Disseny del sistema
- Desenvolupament
- Proves
- Implantació

S'ha escollit aquesta metodologia perquè és la més extensa en l'àmbit del desenvolupament de sistemes d'informació en el món empresarial i és on tinc més experiència, punt transcendental si has de decidir entre diferents estratègies.

1.4 Planificació del Treball

Per aconseguir a realitzar d'una manera adequada aquest projecte s'ha decidit fer-ho per fases o etapes. Aquestes fases o etapes són cada una de les PACs mes la memòria final i la presentació de la solució informàtica o producte. Aquests documents han de ser entregats en unes dates determinades o fites.

1.4.1 Planificació Global

Les fases, tasques i dates del treball final de carrera (TFC) són les següents:

FASE	TASCA	INICI	FINAL
PAC1	Pla de treball i anàlisi preliminar	19/09/2015	01/10/2015
PAC2	Anàlisi requeriments i disseny conceptual	02/10/2015	29/10/2015
PAC3	Implementació	30/11/2015	17/12/2015
PAC4	Memòria i Presentació Virtual	18/12/2015	12/01/2016
DEBAT	Defensa del TFC	25/01/2016	28/01/2016

Taula 1. Planificació global

1.4.2 Planificació Especifica

A continuació detallem el pla de treball que es mostra al quadre anterior:

ID	Tasques	Durada	Hores	Data Inici	Data Fi
1	**Inici del TFC**	**134**	**301,25**	**17/09/2015**	**28/01/2015**
2	Presentació al fòrum	1	0,25	17/09/2015	17/09/2015
3	Lectura del Pla docent	1	1	18/09/2015	18/09/2015
4	**Inici PAC1**	**13**	**47,25**	**19/09/2015**	**01/10/2015**
5	Lectura Anàlisi PAC1	2	8	19/09/2015	20/09/2015
6	Estudi Data warehouse Magatzems de dades i models multidimensionals	3	9	21/09/2015	23/09/2015
7	Elaboració Anàlisi Requeriments	4	14	24/09/2015	27/09/2015
8	Elaboració Pla treball	4	16	28/09/2015	01/10/2015
9	Lliurament PAC1	0	0,25	01/10/2015	01/10/2015
10	**Inici PAC2**	**28**	**67,25**	**02/10/2015**	**29/10/2015**
11	Lectura Anàlisi PAC2	2	8	02/10/2015	04/10/2015
12	Estudi Material didàctic o teòric	5	7	05/10/2015	09/10/2015
13	Anàlisi detallat de requeriments	11	30	10/10/2015	20/10/2015
14	Disseny del model dimensional	9	22	21/10/2015	29/10/2015
15	Lliurament PAC2	0	0,25	29/10/2015	29/10/2015
16	**Inici PAC3**	**49**	**123,25**	**30/10/2015**	**17/12/2015**
17	Lectura Anàlisi PAC3	2	8	30/10/2015	31/10/2015
18	Estudi Material didàctic o teòric	5	7	01/11/2015	05/11/2015

19	Construcció magatzem de dades	12	32	06/11/2015	17/11/2015
20	Configuració eina explotació de dades	6	24	18/11/2015	27/11/2015
21	Construcció Informes	10	28	28/11/2015	07/12/2015
22	Anàlisi de la informació	4	8	08/12/2015	11/12/2015
23	Fase de proves i validacions	6	16	12/12/2015	17/12/2015
24	Lliurament PAC3	0	0,25	17/12/2015	17/12/2015
25	**Inici PAC4**	**26**	**60,25**	**18/12/2015**	**12/01/2016**
26	Lectura Anàlisi PAC4	2	8	18/12/2015	19/12/2015
27	Estudi Material didàctic o teòric	2	6	20/12/2015	21/12/2015
28	Elaboració memòria	12	20	22/12/2015	02/01/2016
29	Elaboració presentació virtual	10	26	03/01/2016	12/01/2016
30	Lliurament PAC4	0	0,25	12/01/2016	12/01/2016
31	Debat i defensa	4	2	25/01/2016	28/01/2016

Taula 2. Planificació especifica

1.4.3 Descripció de les tasques

(ID 2) Presentació al fòrum: Cada alumne es presenta al fòrum de l'aula mitjançant un correu electrònic.

(ID 3) Lectura del Pla docent: Revisar el pla docent de l'assignatura, en aquest cas el TFC, per a estar assabentat de les normes o regles del projecte.

(ID 5) Lectura Anàlisi PAC1: Llegir la PAC publicada detingudament i analitzar tots els punts per tenir-ho tot clar abans de començar.

(ID 6) Estudi Data Warehouse Magatzems de dades i models multidimensionals: Lectura teòrica i estudi del material per adquirir coneixements que seran necessaris per desenvolupar la PAC i el TFC en general.

(ID 7) Elaboració Anàlisi Requeriments: Document formal on es descriu l'estudi del cas del TFC. Aquest és fonamental abans de qualsevol implementació, ja que s'estudien les fonts de dades i s'identifiquen els diferents elements del DW. Feina necessària per realitzar els informes demanats.

(ID 8) Elaboració Pla treball: Document formal on es descriu per una part les diferents fases del projecte amb la seva estimació de dates corresponent i per una altra part es descriu el maquinari i programari necessari a més dels possibles riscos que posen en perill el projecte.

(ID 11) Lectura Anàlisi PAC2: Llegir la PAC publicada detingudament i analitzar tots els punts per tenir-ho tot clar abans de començar.

(ID 12) Estudi Material didàctic o teòric: Lectura teòrica i estudi del material per adquirir coneixements que seran necessaris per desenvolupar la PAC i el TFC en general.

(ID 13) Anàlisi detallat de requeriments: Document formal amb una anàlisi acurat de les fonts de dades i els requeriments.

(ID 14) Disseny del model dimensional: Document formal amb el disseny de tot el projecte: arquitectures, model multidimensional, disseny físic, descripció del procés ETL a alt nivell (pseudocodi), etc.

(ID 17) Lectura Anàlisi PAC3: Llegir la PAC publicada detingudament i analitzar tots els punts per tenir-ho tot clar abans de començar.

(ID 18) Estudi Material didàctic o teòric: Lectura teòrica i estudi del material per adquirir coneixements que seran necessaris per desenvolupar la PAC i el TFC en general. En aquest cas es tracta d'estudiar els documents d'ajuda del programari a utilitzar.

(ID 19) Construcció magatzem de dades: Creació de la base de dades en la fase d'implementació.

(ID 20) Configuració eina d'explotació de dades: Construcció i configuració del procés ETL, modelatge i planificació de càrrega de dades.

(ID 21) Construcció Informes: Elaboració dels informes demanats en el projecte.

(ID 22) Anàlisi de la informació: Verificació de les dades dins el flux de dades.

(ID 23) Fase de proves i validacions: Creació d'un joc de proves per testejar totes les casuístiques possibles amb la finalitat de trobar errors i comprovar la consistència del model.

(ID 26) Lectura Anàlisi PAC4: Llegir la PAC publicada detingudament i analitzar tots els punts per tenir-ho tot clar abans de començar.

(ID 27) Estudi Material didàctic o teòric: Lectura teòrica i estudi del material per adquirir coneixements que seran necessaris per desenvolupar la PAC i el TFC en general. En aquest cas es tracta d'estudiar els documents d'ajuda referents a presentacions de documents i elaboració de presentacions.

(ID 28) Elaboració memòria: Fer el document formal final, que és un compendi de totes les pacs anteriors.

(ID 29) Elaboració presentació virtual: Preparar una presentació multimèdia de la solució del projecte.

(ID 31) Debat i defensa: Defensa del projecte davant un tribunal de consultors de la UOC.

1.4.4 Diagrama de *Gannt*

En la figura 1 podem veure la planificació detallada anterior de manera gràfica per mitjà del programa *GanntProject*.

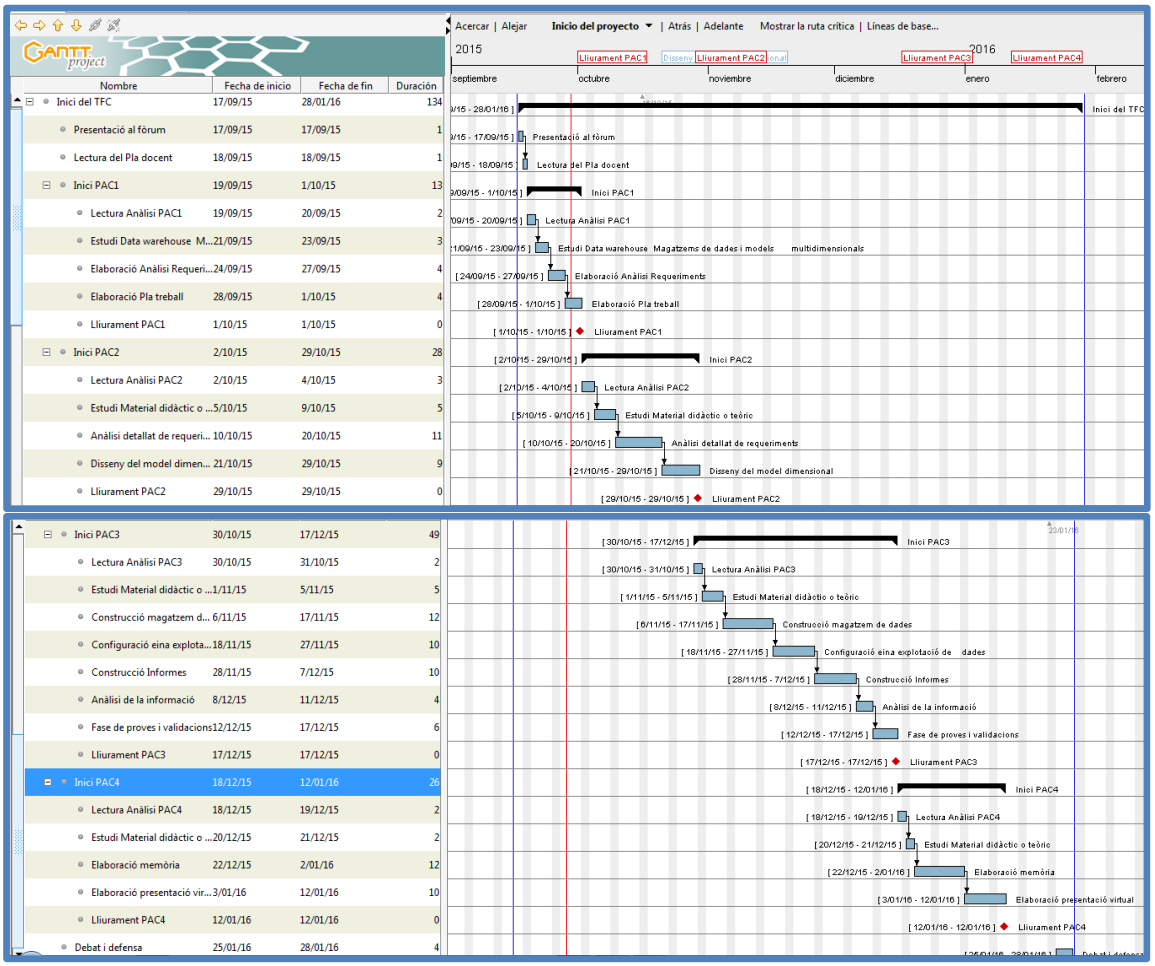


Figura 1. Diagrama de Gannt

1.5 Breu sumari de productes obtinguts

Els productes obtinguts una vegada realitzat el projecte són un magatzem de dades desenvolupat íntegrament en *MySQL* que conté una base de dades relacional on emmagatzemem les dades i un model multidimensional OLAP per explotar els informes.

Els cubs han estat desenvolupats amb *schema-workbench* de *Mondrian*. Per carregar les dades al nostre espai d'emmagatzematge i posteriorment al model multidimensional hem utilitzat *Spoon* de *Pentaho*.

Finalment tenim una llista d'informes usant el *plugin Saiku* de *BA Pentaho*.

A continuació veiem la llista de productes que es troben a la màquina d'Amazon:

Fitxer de base de dades i model multidimensional: **TFC.mwb**

Transformacions i Jobs de procés ETL, sota la nomenclatura:

- JOB_INIT*.kjb
- JOB_DELTA*.kjb
- TRANS_INIT_*.ktr
- TRANS_DELTA_*.ktr

Model cub *star-schema*: **Erasmus_Cube.xml**

Els informes estan publicats al servidor de *Pentaho*: **localhost:8080/pentaho/**

1.6 Breu descripció dels altres capítols de la memòria

Capítol 2. *Requeriments*. Etapa inicial en el desenvolupament del magatzem de dades on recollim tots els requisits funcionals i no funcionals de la solució.

Capítol 3. *Arquitectura (Maquinari i Programari)*. Breu descripció tant del maquinari com del programari utilitzat per desenvolupar el magatzem de dades i posteriors informes.

Capítol 4. *Base de dades Storage Mobility*. Detall de la base de dades relacional prèvia al model multidimensional, on trobem les taules Mobilitat1, 2, 3 i les taules de dades mestres.

Capítol 5. *Model Conceptual*. Definició del model en estrella de mobilitat, taula de fets taules de dimensions, mesures, atributs.... així com el diagrama del model.

Capítol 6. *Model Multidimensional*. Explicació de com s'ha construït el cub utilitzant el programa *schema-workbench Mondrian.*

Capítol 7. *Càrrega de dades inicial*. Procés ETL de càrrega de dades inicial a partir dels fitxers proporcionats. S'explica com s'han transformat les dades fins a deixar-les al cub.

Capítol 8. *Càrrega de dades delta*. Procés ETL per a càrregues vinents, és a dir, càrregues de cursos futurs i de dades mestres futures.

Capítol 9. *Gestió d'errors*. Fase on es mostra com es tracten els errors en el procés ETL, i com es poden consultar per a intentar corregir-los.

Capítol 10. *Informes*. Presentació dels informes finals, construïts en *Saiku*, després de fer la càrrega de la base de dades.

Capítol 11. *Suposicions*. Explicació dels diferents punts del projecte que necessiten matisacions per a la seva bona entesa i evitar confusions.

Capítol 12. *Conclusions*. Part final de la memòria on es fan les reflexions adequades de tot el projecte.

Capítol 13. *Glossari*. Llistat de conceptes i acrònims, amb la seva definició, utilitzats en aquesta memòria.

Capítol 14. *Annexos*. Podem trobar l'enunciat del TFC.

Capítol 15. *Bibliografia*. Llista de les referències utilitzades, siguin libres o URL web, que s'han utilitzat com a suport en la realització de la solució.

2. REQUERIMENTS

Els requeriments es poden classificar en dos grans grups: els que fan referència a les necessitats que ha de satisfer el sistema (què ha de fer) i els que expressen restriccions sobre el conjunt de solucions possibles (com ho ha de fer). Dels del primer grup en diem requisits funcionals, mentre que dels del segon grup en diem requisits no funcionals. Els requeriments funcionals fan referència a la funcionalitat que ha de proporcionar el sistema i ens indiquen quin és el comportament del sistema davant dels estímuls que li arriben de l'exterior. Els requeriments no funcionals acostumen a tenir forma de restricció i acostumen a afectar gran part del sistema, sense incloure-hi comportament.

2.1. Requeriments funcionals

La funcionalitat que ha de proporcionar el sistema és bàsicament mostrar informació del magatzem de dades als usuaris per mitjà d'informes. En el cas que ens ocupa són informes d'estudiants que cursen assignatures en diferents institucions de diferents països dins el marc del programa d'Erasmus, pel seu posterior anàlisi. Per dur a terme aquesta solució informàtica haurem d'utilitzar el següent:

Un sistema gestor de bases de dades per a construir el magatzem de dades.

Una Suite de BI per la construcció d'informes.

Utilització d'eines ETL (*Extract, Transformation & Load*) per normalitzar o harmonitzar les dades provinents dels fitxers que proporciona Erasmus.

Una eina de modelatge per construir la nostra solució multidimensional.

2.2. Requeriments no funcionals

Erasmus ens ha proporcionat un document amb totes les regles i restriccions que ha de seguir la nostra base de dades, aquest és el "Student Mobility datadictionary".

Per fer el modelatge haurem de fer servir un model multidimensional en forma d'estrella, més conegut com a "Star schema".

El sistema ha d'estar preparat per actualitzacions de càrregues futures.

Tota la informació ha de poder ser agregada per any, de manera que puguem utilitzar tots els informes d'una manera comparativa i/o evolutiva per any.

La UOC ens proporcionarà un espai per fer anar una màquina virtual d'Amazon.

3. ARQUITECTURA (MAQUINARI I PROGRAMARI)

Al món de la informàtica, terme general que s'aplica a l'estructura d'un sistema informàtic o d'una part del mateix. És aplicable al disseny del software de sistema, així com al hardware que utilitza.

3.1. Maquinari

Disposem d'una plataforma remota Amazon EC2 (*Elasitc Compute Cloud*). És una part central de la plataforma de computació al núvol de l'empresa Amazon.com denominada *Amazon Web Services* (AWS). EC2 permet als usuaris rentar computadores virtuals on poder córrer les seves pròpies aplicacions. Aquest tipus de servei suposa un canvi al model informàtic al proporcionar capacitat informàtica amb mida modificable al núvol, pagant per la capacitat utilitzada. En lloc de comprar o llogar un determinat processador para utilitzar-lo mesos o anys, en EC2 es lloga la capacitat per hores. Veure figura 2.

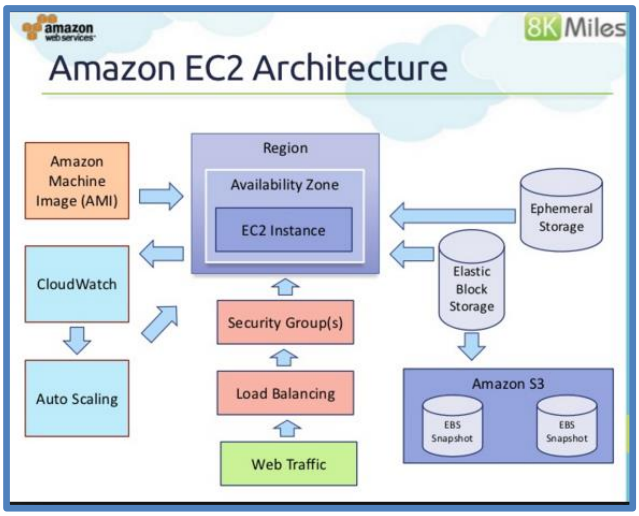


Figura 2. Arquitectura Màquina Amazon

3.2. Programari

La finalitat del projecte és proporcionar uns informes de mobilitat d'estudiants d'Erasmus. Per aconseguir aquest objectiu necessitem un conjunt d'eines, amb una funcionalitat especifica cadascuna d'elles.

El sistema operatiu on estaran instal·lades totes les eines és Lubuntu versió 14.04.3 LTS amb les característiques que es mostren a la figura 3 i 4:

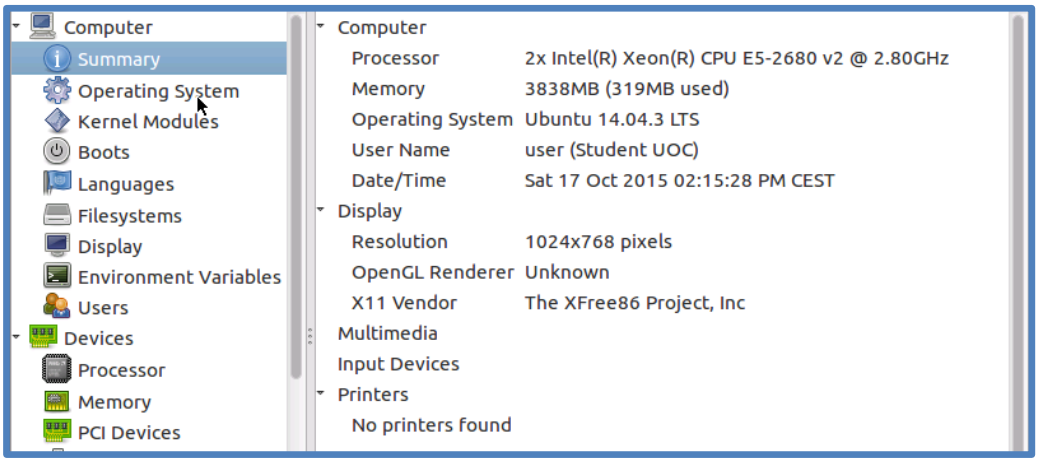


Figura 3. Propietats PC virtual a Amazon

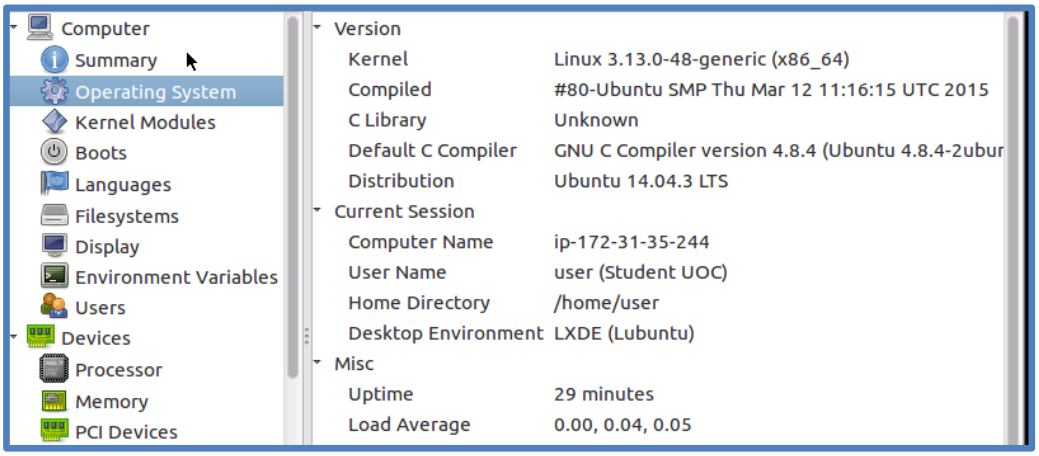


Figura 4. Propietats Sistema Operatiu màquina virtual a Amazon

Dins aquest sistema operatiu trobem:

- **Mysql:** Servidor SQL utilitzat.
- **Mysql-workbench versió 6.3 d'Oracle:** Eina de modelatge de base de dades que utilitzarem per crear la base de dades d'Erasmus.
- **Pentaho BA versió 5.4:** Entrem al món de Pentaho, necessari per explotar les dades que estan emmagatzemades a mysql. Disposem de diverses aplicacions dins l'univers Pentaho.
- **PDI (Kettle Spoon) versió 5.4:** Entorn gràfic que utilitzarem en el procés d'extracció de les dades, transformació i càrrega de dades. Necessari per depurar les dades.
- **Mondrian Schema-Workbench versió 3.10:** Entorn gràfic pel modelatge de cubs (model multidimensional).
- **Saiku versió 5.4:** Finalment amb aquesta eina donarem forma als informes a partir de les dades emmagatzemades al cub.

4. BASE DE DADES *STORAGE MOBILITY*

Després de veure en detall les fonts de dades, hem de pensar, ¿on estaran les dades abans d'arribar al model multidimensional? La idea és tenir una primera capa volàtil (S'esborren cada vegada que carregues) on carregarem totes les dades dels fitxers transaccionals a la base de dades.

Per fer això crearem 2 taules (Mobilitat 1 i Mobilitat 2), la primera pel curs 2011/12 i la segona pel curs 2012/13. El motiu de fer aquesta primera capa és perquè les dues fonts de dades tenen estructures diferents i necessiten un tractament abans de carregar les dades a la segona capa. En futures càrregues de dades, aquestes entraran directament a la taula corresponent al format del Curs 2012/13.

Les dades mestres es carregaran directament des de fitxer a tres taules que seran: Assignatures, Països i Institucions. Paral·lelament crearem unes taules de dades mestres per: estudiants, empreses, cursos, temps i tipus de mobilitat que es nodriran de les dades transaccionals dels dos fitxers, generant automàticament un codi unívoc en el moment de la càrrega.

En la segona capa integrarem les dades de Mobilitat 1 i Mobilitat 2 en una tercera taula (Mobilitat 3) que contindrà tots els camps de les dues anteriors més un camp nou (Curs). Aquesta taula l'omplirem amb l'ajuda de la taula de dades mestres d'estudiants per assignar codis nous. D'aquesta manera evitem la manca de dades d'estudiants que tenim en el fitxer del curs 2011/12. Per altra banda, amb el camp nou de Curs podem diferenciar les dades de la seva procedència sense haver de mirar les dates de tots els registres.

Al final el dibuix quedaria de la següent manera:

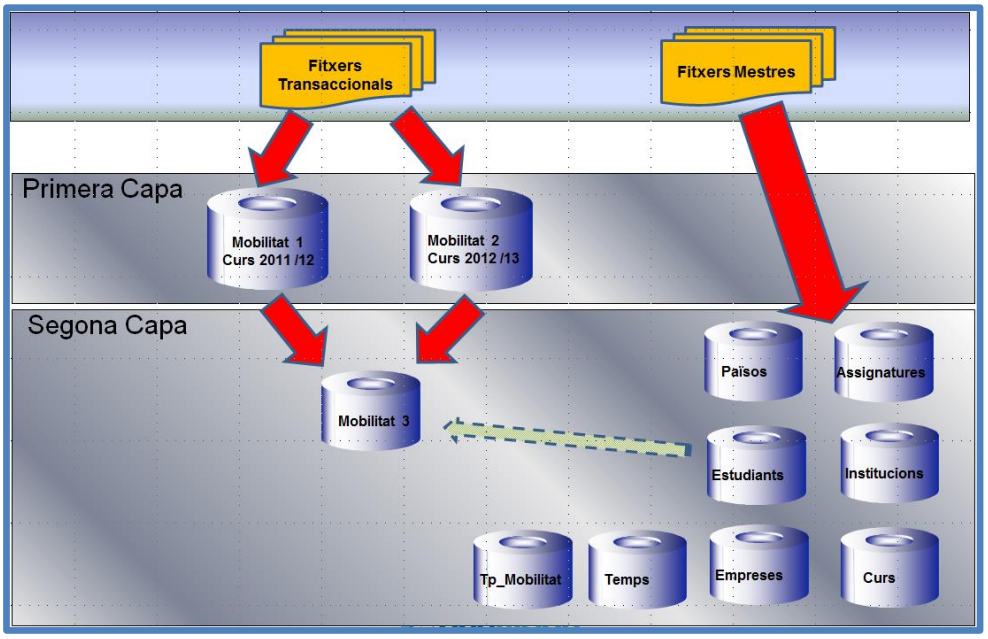


Figura 5. *Storage Mobility*

5. MODEL CONCEPTUAL

En aquest punt hem de fer el disseny conceptual del model multidimensional. Bàsicament veurem dos grans grups, com són la tria del fet i el corresponent conjunt de dimensions. Aquesta particular estructura és anomenada *star-schema* (Estrella), per la manera com es representa el disseny, com es pot veure en l'exemple següent:

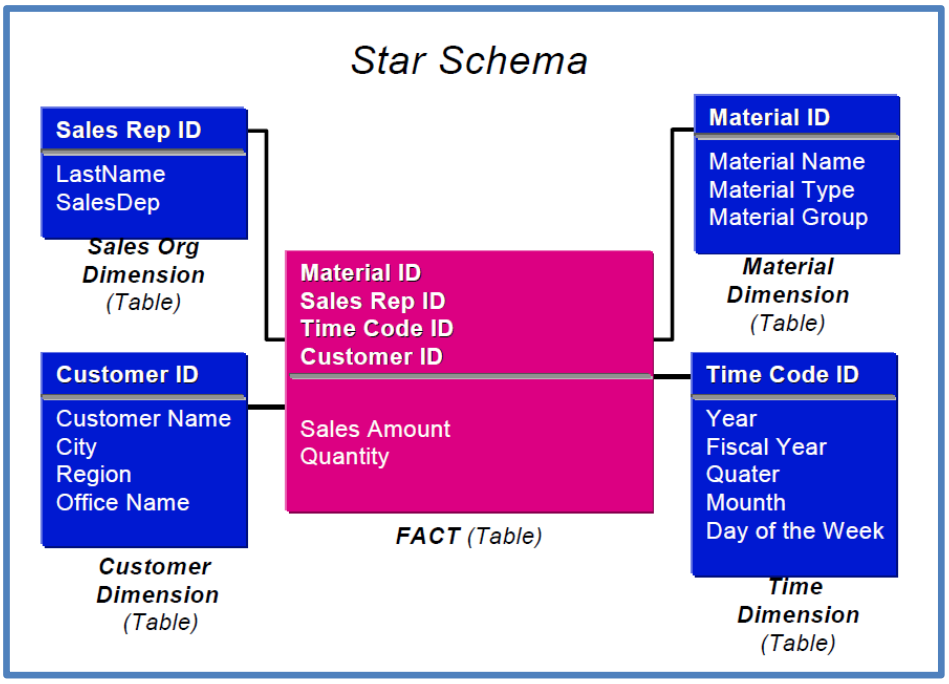


Figura 6. Diagrama esquema estrella (Star-schema)

5.1. Tria del fet

En una base de dades la taula de fets és la peça central de l'esquema multidimensional i conté els valors o indicadors del model que s'estudia. Podríem dir que és la part més important de l'estrella.

En el cas que ens ocupa, és la mobilitat d'estudiants d'Erasmus entre els diferents països, per tant la nostra taula de fets tindrà els identificadors d'estudiants, assignatures cursades, emplaçaments, països d'institucions així com de l'estudiant, el tipus de mobilitat i el curs realitzat.

Entenem com emplaçaments les diferents localitzacions on l'estudiant cursa les assignatures, sigui institució o empresa, i les institucions de procedència.

Com veurem a continuació, a part de la taula de fets tenim 8 taules de dimensions el que fa que el nostre model sigui de 8 dimensions.

5.2. Tria del grànul escaient

El grànul defineix el nivell de detall més baix que tindrà el nostre model multidimensional. Aquest punt també és important, ja que si el nivell de detall (grànul fi) és molt baix, més gran serà el nombre de registres a la taula de fets.

En el nostre cas l'estudiant seria el nivell de detall més baix al qual podríem accedir.

5.3. Tria de les dimensions

En qualsevol magatzem de dades, la construcció d'un model multidimensional requereix una taula de fets, comentada anteriorment, i diverses taules de dimensions, aquestes determinen els paràmetres que depenen dels fets registrats a la taula de fets. Podríem dir que la taula de fets conté les dades d'interès i les taules de dimensions contenen metadades sobre aquests fets.

Al model de dades de mobilitat d'Erasmus tenim 8 dimensions que definim a continuació:

- **Assignatures:** conjunts d'àrees o assignatures que un estudiant pot cursar en qualsevol de les institucions repartides al continent.
- **Curs:** llista de cursos existents a la base de dades.
- **Empresa:** empreses on els estudiants d'Erasmus realitzen els seus estudis pràctics.
- **Estudiant:** aquí emmagatzemem les dades rellevants dels estudiants.
- **Institucions:** conjunt d'institucions, universitats emissores i receptores d'estudiants que participen en el programa d'Erasmus. Com tenim diferents dates la considerem del tipus *role-playing dimension.*
- **Països:** llistat de països que participen en el programa Erasmus. Com tenim diferents dates la considerem del tipus *role-playing dimension*.
- **Temps:** aquesta dimensió està present en pràcticament tots els models multidimensionals d'estrella. Pel model que estem definint, la necessitem per saber quan un estudiant s'ha incorporat a un curs. Com tenim diferents dates la considerem del tipus *role-playing dimension*.
- **Tipus de Mobilitat:** dimensió que conté els tres tipus de mobilitat (S, C o P) que existeixen actualment al programa Erasmus.

5.4. Tria dels atributs de cada dimensió

Ara que ja estan definides les dimensions, és el moment d'establir quins seran els seus atributs. S'ha intentat que el nom dels atributs siguin prou entenedors i descriptius. Els atributs que en resulten per les diferents dimensions són els que presento tot seguit:

Dimensió Assignatures. Aquesta dimensió conté totes les àrees / assignatures que cursen els estudiants d'Erasmus. Donat que només tenim un atribut que és el mateix identificador, aquesta dimensió passa a ser part de la taula de fets com una columna més, és el que diem una dimensió degenerada.

- **pkassignatura:** Clau primària, de tipus *integer* i auto incrementable.
- **fsourcecode:** De tipus *string* de longitud 4 i clau primària. L'assignatura estudiada per l'estudiant a la institució emissora. L'identificador pot pertànyer a la codificació ISCED97 o a la codificació d'Erasmus.
- **Fdesc:** De tipus *string* de longitud 75. Descripció de l'assignatura cursada.

Dimensió Curs. Aquesta dimensió és un contenidor de cursos que es van carregant a la base de dades d'Erasmus.

- **pkCurs:** clau primària de tipus *integer* i auto incrementable.
- **Curs:** De tipus *string* de longitud 7. Nom del curs de format AnyInici "/" AnyFi en forma AAAA + "/" + AA.

Dimensió Empresa. Aquesta dimensió fa referència als diferents emplaçaments que entren en joc al programa d'Erasmus. En aquest cas estem parlant del món de les empreses que col·laboren amb el programa. Vegem quins són els atributs:

- **IdEmpresa:** Clau primària, de tipus *integer*. Codi de l'empresa / organització on l'estudiant fa les seves pràctiques.
- **fdesc:** De tipus *string* de longitud 255. Nom de l'empresa / organització a on un estudiant ha fet les seves pràctiques.

Dimensió Estudiant. Aquesta dimensió s'utilitzarà per identificar tots els estudiants que participen en el programa d'Erasmus al llarg dels anys. Els atributs són els següents:

- **pkEstudiant:** De tipus *integer* i clau primària.
- **Edat:** De tipus *integer*. Són els anys que té cada estudiant.
- **Sexe:** De tipus *string* i longitud d'1. Identifica si un estudiant es Home o dona.
- **Identificador:** De tipus *string* de longitud 20. Codi alfanumèric, identificador d'estudiant, semblant al DNI.

Dimensió Institucions. Aquesta dimensió fa referència als diferents emplaçaments que entren en joc al programa d'Erasmus, ja siguin emissors com receptors. Vegem quins són els atributs:

- **Pkinstitucions:** clau primària de tipus *Integer* i auto incrementable.
- **fsourcecode:** De tipus *string* de longitud 16. Codi de la institució on l'estudiant cursa el seu període d'Erasmus.
- **fchartercode:** De tipus *string* de longitud 4.
- **fdesc:** De tipus *string* de longitud 130. Nom de la universitat que participa a Erasmus.
- **fcarrer:** De tipus *string* de longitud 100. Nom del carrer de la universitat.

- **fcp:** De tipus *string* de longitud 40. Codi postal de la universitat.
- **fciutat:** De tipus *string* de longitud 60. Nom de la ciutat on es troba la universitat.

Dimensió Països. Aquesta dimensió conté la llista de països que participen en el programa Erasmus. Vegem quins són els atributs:

- **pkpais:** Clau primària, de tipus *Integer* i auto incrementable.
- **fsourcecode:** De tipus *string* de longitud 4. Codi de país de la llista de participants.
- **fdesc:** De tipus *string* de longitud 255. Nom del país que participa a Erasmus.

Dimensió Temps. Aquesta dimensió és bàsica per qualsevol model, atès que el temps sempre és una de les característiques requerides per analitzar la informació de qualsevol informe. Els atributs són els següents:

- **pkMesAny:** De tipus *integer* i clau primària. Identifica el moment del temps on l'estudiant ha cursat els seus estudis.
- **Any:** De tipus *integer*, any d'estudis. Es dóna el cas que un any pot pertànyer a dos cursos diferents.

Dimensió Tipus Mobilitat. Aquesta dimensió contindrà els 3 tipus de mobilitat d'Erasmus (S, C o P).

- ***S: Student mobility only* –** L'estudiant surt d'una institució emissora i va a una institució receptora. No hi ha pràctiques a empresa.
- ***C: Study mobility combined with a placement* –** L'estudiant combina l'estudi a una institució receptora amb pràctiques a empresa.
- ***P: Placement mobility:*** L'estudiant només es trasllada per fer pràctiques a una empresa.

Atributs:

  - **pkMobilitat:** De tipus *integer* i clau primària.
  - **Tipus:** De tipus *string* de longitud 1. Codi que identifica el tipus de mobilitat que realitza l'estudiant (S, C o P).
  - **Desc:** Descripció del tipus de mobilitat. De tipus *string* de 45.

5.5. Tria dels atributs de la taula de fets

La taula de fets és un conglomerat d'identificadors, els quals estan relacionats unívocament amb les taules de dimensió. A part tenim les mesures o indicadors del negoci o del projecte que s'estigui realitzant. A partir d'aquí, donat que ja hem vist les dimensions en detall, és moment ara de veure la taula de fets:

Taula de fets. Taula de mobilitat d'estudiants entre les universitats i empreses adscrites al programa Erasmus. Vegem els atributs:

- **pkassignatura:** Identificador de dimensió d'assignatures.
- **pkStudiant:** Identificador de dimensió d'estudiant.
- **pkpaisSt:** Identificador de dimensió països referent a l'estudiant.
- **pkInstiE:** Identificador de dimensió d'institució emissora.
- **pkpaisIE:** Identificador de dimensió països referent a la institució emissora.
- **pkInstiR:** Identificador de dimensió d'institució receptora.
- **pkpaisIR:** Identificador de dimensió països referent a la institució receptora.
- **pkEmpresa:** Identificador de dimensió d'empreses.
- **pkpaisIR:** Identificador de dimensió països referent a l'empresa.
- **IdInstitucioR:** Identificador de dimensió d'institució receptora.
- **pkMesAnyI:** Identificador de dimensió de temps referent a institucions.
- **pkMesAnyIE:** Identificador de dimensió de temps referent a empreses.
- **pkTMob:** Identificador de dimensió de tipus de mobilitat.
- **pkCurs:** Identificador de dimensió de Curs.
- **Grant:** Quantitat en euros de les beques.
- **Cont:** Comptador de registres.
- **Edat:** Edat d'estudiants.

5.6. Distingir entre descriptors i jerarquies d'agregació

D'entre els atributs que hi ha en una dimensió, n'hem de distingir dos tipus: els que utilitzarem per a agrupar i els que serviran simplement per a seleccionar.

Descriptors: MesAny, Any, Curs, Àrea, Edat, Sexe, Nacionalitat, Institució emissora, país emissor, Institució receptora, país receptor, tipus de mobilitat.

Jerarquies: MesAny, Any.

S'ha de dir que mes i any a part de ser agrupadors també poden servir per a fer seleccions.

5.7. Decidir quines són les mesures que interessen

Les Mesures són atributs numèrics normalment additius. En el nostre projecte tenim les següents mesures:

***Grant*:** Quantitat en euros de les beques per cada estudiant.

Cont: Comptador de registres, de gran utilitat quan es volen fer càlculs tenint en consideració el nombre de registres. La idea és assignar el valor 1 a cada registre, de manera que podrem contar en l'agregació.

Edat: edat de cada estudiant. Sabem que l'edat és un clar atribut de la taula estudiants però com es demana fer una sèrie de càlculs amb l'edat, he considerat adient afegir-lo com una mesura més.

Les dades numèriques que es demanen en els informes (%, Avg, ...) no és necessari emmagatzemar-les, es poden calcular en temps d'execució.

5.8. Definir cel·les

Així com en un primer moment pensàvem que només tindríem una única cel·la, després de la inclusió de la mesura edat podem afirmar que tindrem dos. La primera és la *Grant* de la institució emissora en un curs determinat i de la institució receptora. La segona és l'edat mitjana d'estudiants per nacionalitat receptora i emissora. On l'edat mitjana es pot calcular partint de l'agregació de l'edat en qualsevol nivell, com a base de càlcul i ajudant-nos del comptador de registres.

Els percentatges que es demanen en els altres informes es poden calcular en temps d'execució.

5.9. Diagrama Model conceptual *Data Warehouse*

Un cop definit el model conceptual amb les seves taules de dimensió i la taula de fets, ja només queda per mostrar el dibuix o diagrama, com podem veure a la figura 7.

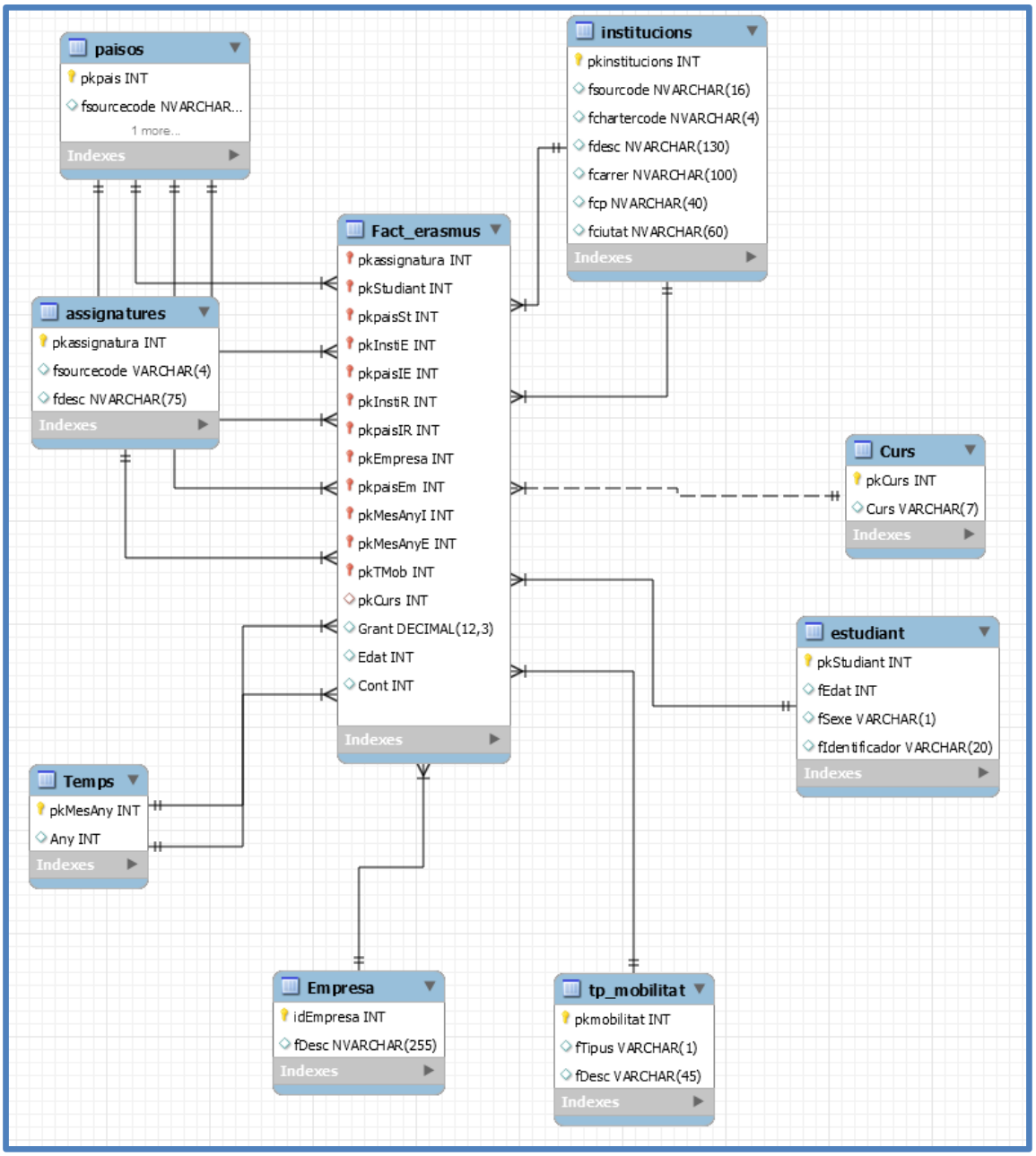


Figura 7. Diagrama model conceptual

6. MODEL MULTIDIMENSIONAL

Finalment i després de l'etapa de disseny mes les posteriors modificacions fetes en la fase d'implementació, el model multidimensional queda com es veu a la figura 8.

Una vegada muntada la BBDD i creades les taules de dimensió i de fets, per construir el cub utilitzem *Schema-workbench* de *Mondrian* per construir el cub. Una vegada fet el cub aquest es guarda en format xml.

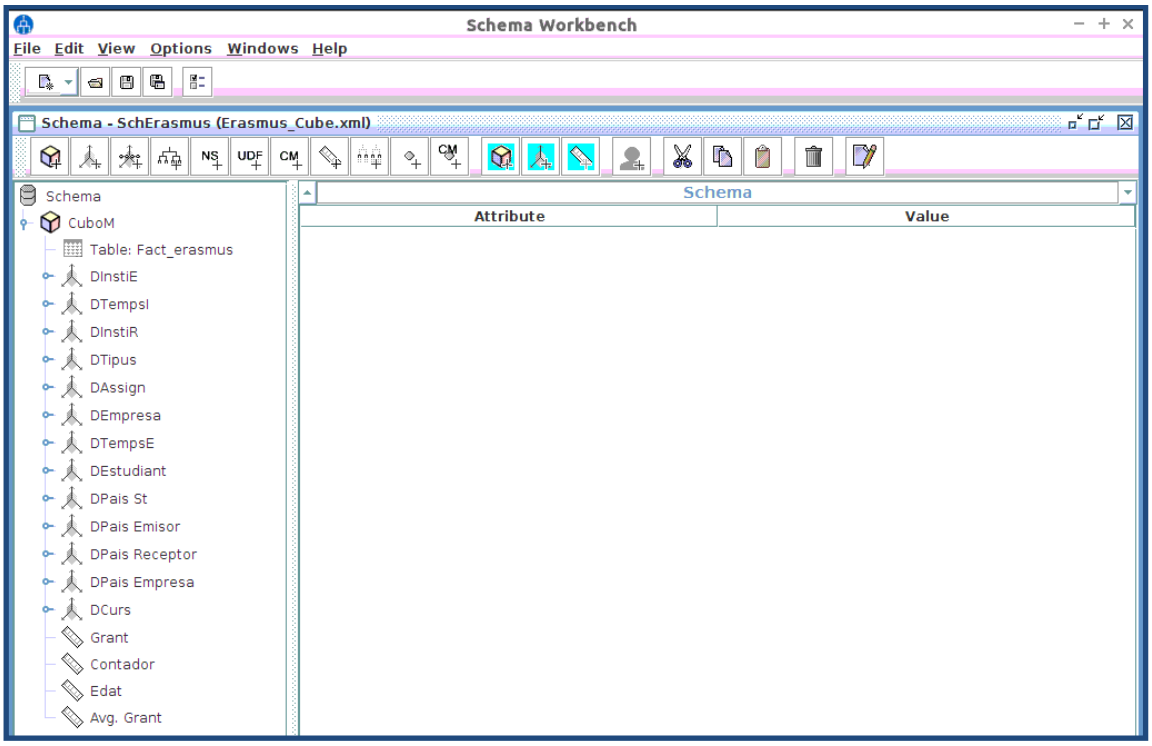


Figura 8. *Modrian* Cub

Com es veu a la figura 8 hi ha més dimensions de les establertes en l'anàlisi inicial, això és perquè amb aquesta eina les dimensions de *role-playing dimension* s'han de fer per separat i per tant has de definir tantes dimensions com camps de relació tinguis a la taula de fets. Només parlem d'aquest tipus de dimensió, les altres, com el cas de la dimensió Curs només en té una.

Per una altra banda s'ha creat una mesura calculada que avalua la mitjana de beques i d'aquesta manera no s'ha de fer a l'informe, com es veu a la figura 9.

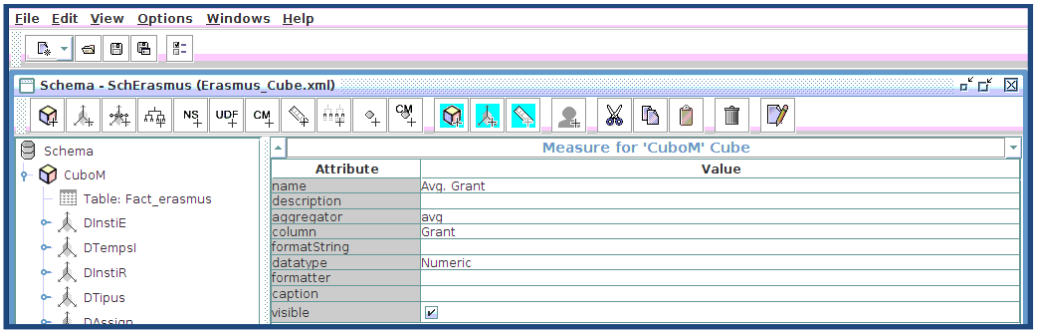


Figura 9. Formula calculada

Al final tenim les dimensions següents a *Schema-workbench* on mostrem l'equivalència a les dimensions reals del format estrella:

DInstiE: Dimensió Institucions

DInstiR: Dimensió Institucions

DTempsI: Dimensió Temps

DTempsE: Dimensió Temps

DTipus: Dimensió Tipus

DAssign: Dimensió Assignatures

DEmpresa: Dimensió Empresa

DEstudiant: Dimensió Estudiant

DPais St: Dimensió País

DPais Emisor: Dimensió País

DPais Receptor: Dimensió País

DPais Empresa: Dimensió País

DCurs: Dimensió Curs

Les mesures *grant*, comptador i edat romanen com s'havien definit a l'etapa d'anàlisi i disseny.

7. CÀRREGA DE DADES INICIAL

La càrrega de dades inicial, l'havíem dividit en dos parts a l'etapa d'anàlisi, una primera part on carregàvem les dades dels fitxers amb scripts *MySQL* i una segona part de tractament i càrrega amb *Spoon*. Això ha canviat, donada la potència del software de *Spoon*, quant a funcionalitats ETL. Finalment tot s'ha fet amb *Spoon*. A continuació veiem totes les transformacions i els diferents *jobs*.

Transformació 1 (TRANS_INIT_Step0.ktr): El primer pas en una càrrega inicial és esborrar les taules, per si ha quedat algun residu del desenvolupament, això ho fem amb un *step* d'SQL, com es veu a la figura 10. Primer de tot hem de treure la restricció de BBDD per esborrar taules amb claus foranes amb la instrucció "SET FOREIGN_KEY_CHECKS = 0;", una vegada hem acabat, ho deixem amb valor 1 com estava.

Figura 10. TRANS_INIT_Step0.ktr

Transformació 2 (TRANS_INIT_Step1.ktr): transformació, com es veu a la figura 11, que afegeix a les taules els codis d'error corresponents, necessaris per si en el procés de càrrega trobem alguna dada incorrecta o desconeguda. Per exemple si ens entra un país no conegut, el sistema li assigna automàticament el codi "*XXXX Unknown*"

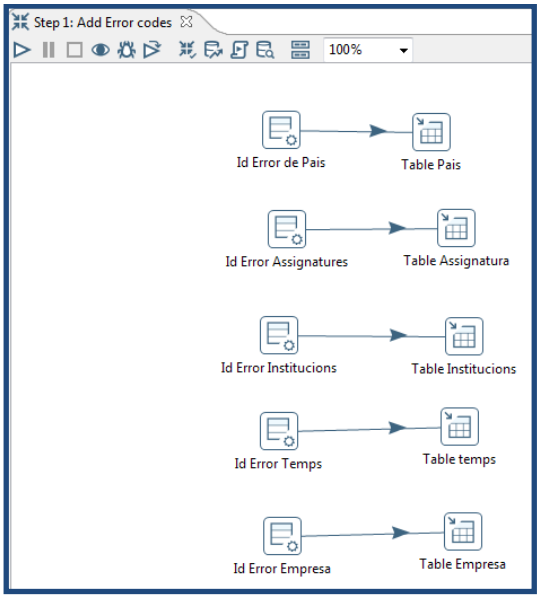


Figura 11. TRANS_INIT_Step1.ktr

Transformació 3 (TRANS_INIT_Step2.ktr): en la següent transformació carreguem les dades mestres de països des del fitxer *.csv. Primer de tot passem el codi de país a majúscules per evitar errors, després validem que els camps siguin correctes respecte a tipus de dades i longitud, ordenem i ens quedem amb els valors únics i llavors gravem a la taula de BBDD de països. En tot moment gestionem els errors, però això ho explicarem més endavant.

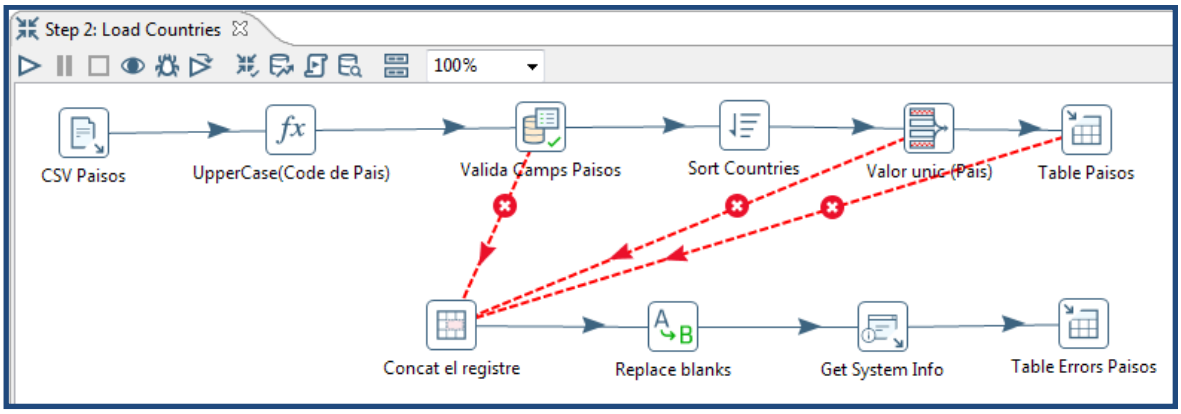


Figura 12. TRANS_INIT_Step2.ktr

Transformació 4 (TRANS_INIT_Step3.ktr): Aquesta transformació carrega les dades mestres d'institucions o universitats tant emissores com receptores. Com veiem a la figura 13 primer de tot fem una validació dels camps respecte a tipus de dades i longitud, després ordenem i ens quedem amb el valor únic. Aquesta seqüència també la seguim per carregar les dades mestres d'assignatures. Finalment carreguem les dades mestres de tipus de mobilitat, que com són 3 valors i no canvien els carreguem directament a la taula de dimensió.

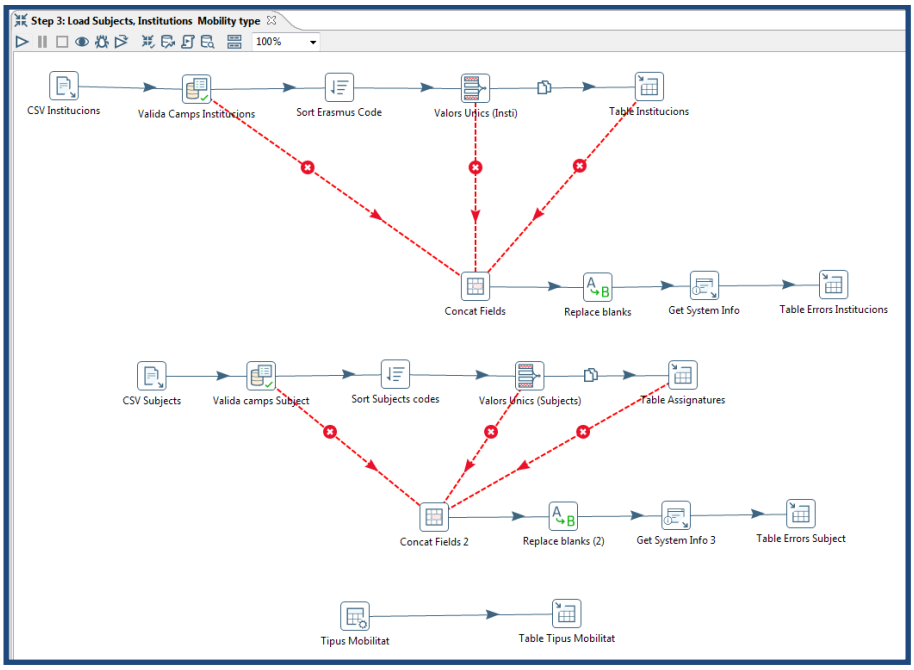


Figura 13. TRANS_INIT_Step3.ktr

Transformació 5 (TRANS_INIT_Step4.ktr): càrrega de les dades transaccionals del Curs 2011/12 a la taula temporal Mobilitat1. Aquí fem una primera purga de les dades, ja que a *Data Validator* mirem el tipus de dades de cada camp i la seva longitud, si és correcte, ja el gravem a la taula.

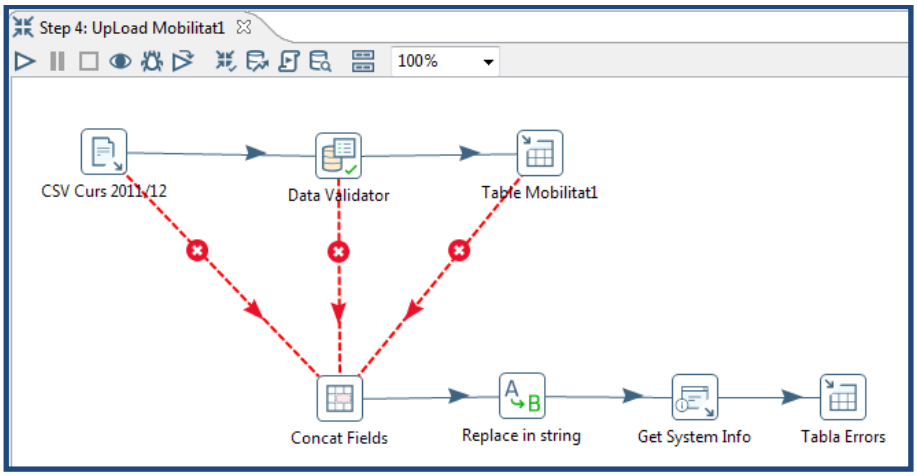


Figura 14. TRANS_INIT_Step4.ktr

Transformació 6 (TRANS_INIT_Step5.ktr): càrrega de les dades transaccionals del Curs 2012/13 a la taula temporal Mobilitat2, veure a la figura 15. Aquí fem una primera purga de les dades, ja que a *Data Validator* mirem el tipus de dades de cada camp i la seva longitud, si és correcte, ja el gravem a la taula, però abans d'això hem de fer un petit tractament al camp "*Placement sector*", a causa de la mala qualitat de dades del fitxer, per això apliquem la següent fórmula per què no falli al següent *step* de *Data Validator*:

"If([TYPE_PLACEMENT_SECTOR_VALUE]=***"?Unknown"***;**""**;[TYPE_PLACEMENT_SECTOR_VALUE])"

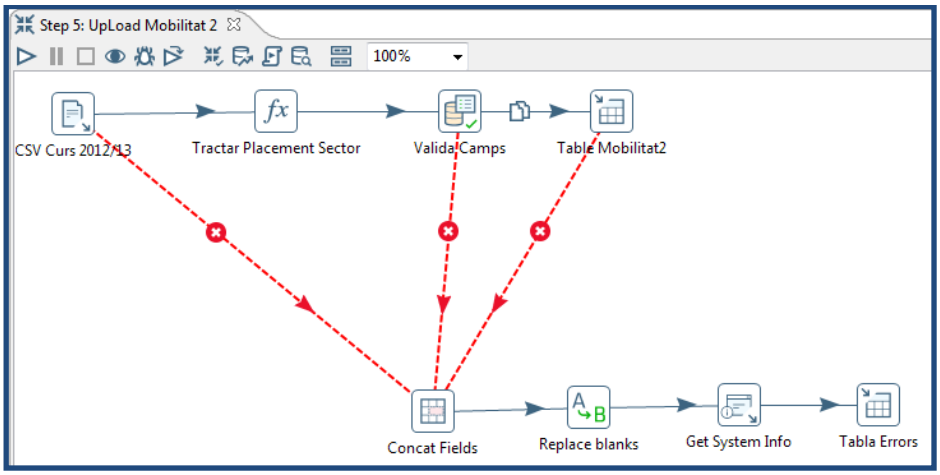


Figura 15. TRANS_INIT_Step5.ktr

Transformació 7 (TRANS_INIT_Step6.ktr): càrrega de dades transaccionals de la taula Mobilitat1 a Mobilitat3, veure a la figura 16. En aquesta transformació fem una segona validació dels camps, però aquesta vegada avaluant el contingut dels camps, tal que s'adaptin a les regles expressades al document: " *Student mobility_datadictionary.pdf*". A part afegim 2 constants el curs, que en aquest cas és 2011/12 i un comptador, el qual en servirà en el futur per fer càlculs.

Aquí hem de comentar que s'ha hagut d'eliminar la validació del camp "*Short duration*" que només permet els valors "X" o "T", perquè sinó no entraven la gran majoria dels registres.

També gravem la dada mestre de Curs ("2011/12"), només en aquest cas, que és el primer, la resta es calcularà de forma dinàmica.

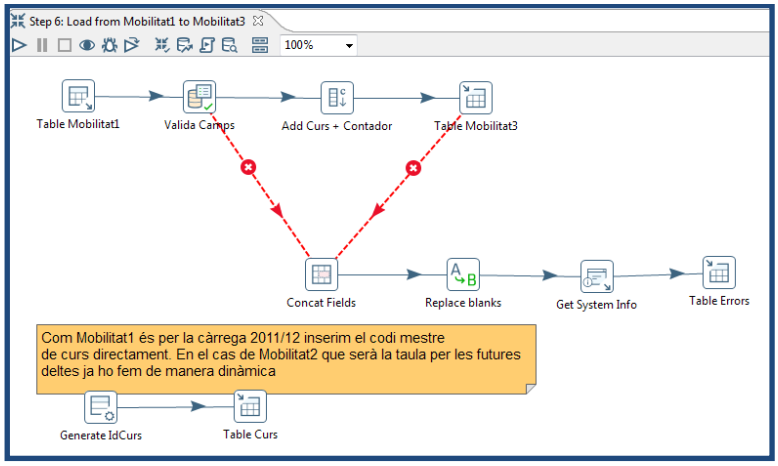


Figura 16. TRANS_INIT_Step6.ktr

Transformació 8 (TRANS_INIT_Step7.ktr): forma dinàmica de carregar el curs. Per això avaluem totes les dades existents al sistema, ajuntant els camps "*fstart_date*" i "*fstart_date_place*". Agafem tots els registres i generem un camp més ("Any"), que omplim retallant les dates dels registres anteriors. A partir d'aquí ens quedem amb els valors únics d'any i els concatenem per muntar el nou valor de Curs que afegim a la taula de dades mestres de Curs.

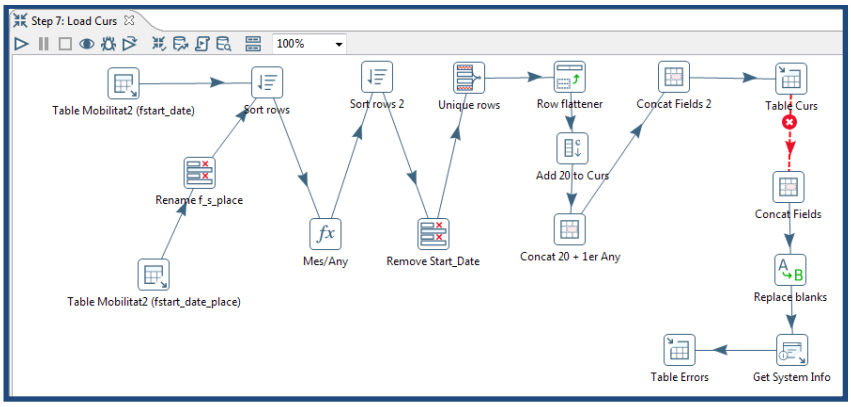


Figura 17. TRANS_INIT_Step7.ktr

Transformació 9 (TRANS_INIT_Step9.ktr): càrrega de dades transaccionals de la taula Mobilitat2 a Mobilitat3, veure a la figura 18. En aquesta transformació fem una segona validació dels camps, però aquesta vegada avaluant el contingut, per tal que s'adaptin a les regles expressades al document: " *Student mobility_datadictionary.pdf*". A part afegim un comptador, el qual en servirà en el futur per fer càlculs.

Aquí hem de comentar que s'ha hagut d'eliminar la validació del camp "*Short duration*" que només permet els valors "X" o "T", perquè sinó no entraven la gran majoria dels registres.

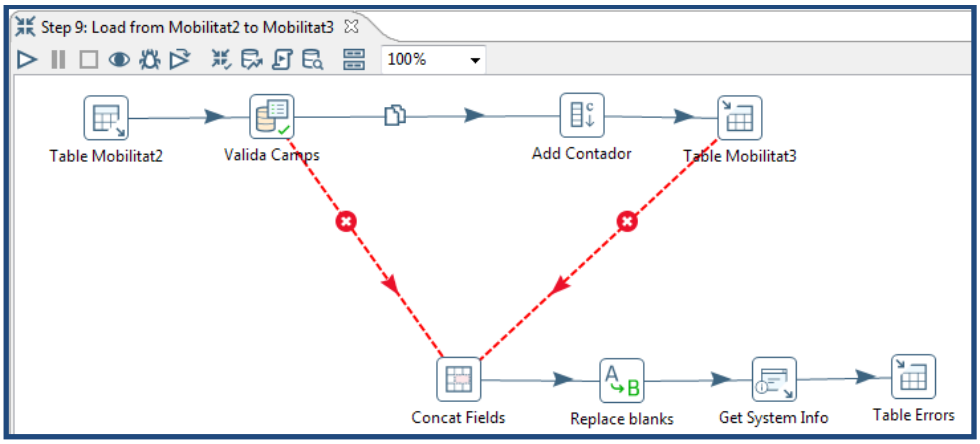


Figura 18. TRANS_INIT_Step9.ktr

Transformació 10 (TRANS_INIT_Step8.ktr): actualitzem les dades en blanc de curs, de la taula Mobilitat3. Prèviament s'han carregat les dades del curs 2012/13 sense informar el curs, per tant és relativament senzill que el sistema identifiqui tots aquells valors que tenen "*Null*" al camp curs.

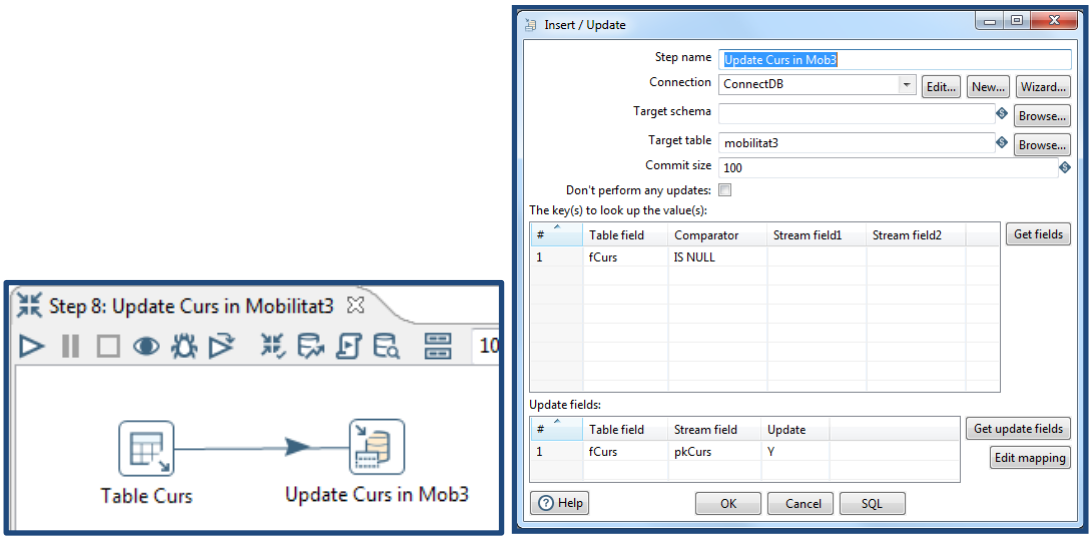


Figura 19. TRANS_INIT_Step8.ktr

Transformació 11 (TRANS_INIT_Step10.ktr): càrrega de dades mestres d'estudiants a partir de la taula de dades transaccionals Mobilitat3, veure a la figura 20. Això ho hem de fer perquè no tenim dades mestres d'estudiants, i per crear-les ho hem de fer considerant que cada registre transaccional és un alumne.

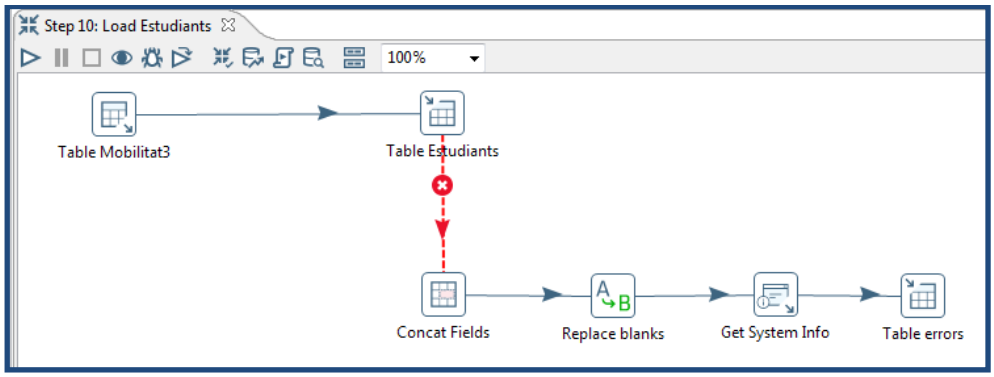


Figura 20. TRANS_INIT_Step10.ktr

Transformació 12 (TRANS_INIT_Step11.ktr): càrrega de dades mestres de temps a partir de la taula de dades transaccionals Mobilitat3, veure a la figura 21. Això ho fem perquè aquesta és una bona manera de muntar les dades mestres de temps. Per fer això ajuntem els dos camps de dates que tenim ("*fstart_date*" i "*fstart_date_place*"), després capturem per separat el mes i l'any. Donat que el camp mes està format amb lletres, "*Aug, Sept...*", utilitzem un *step* de Java amb un codi que fa la transformació a numèric. Una vegada tenim el format YYYYMM, ens quedem amb els valors únics i els gravem a la taula com a *integer* i clau primària.

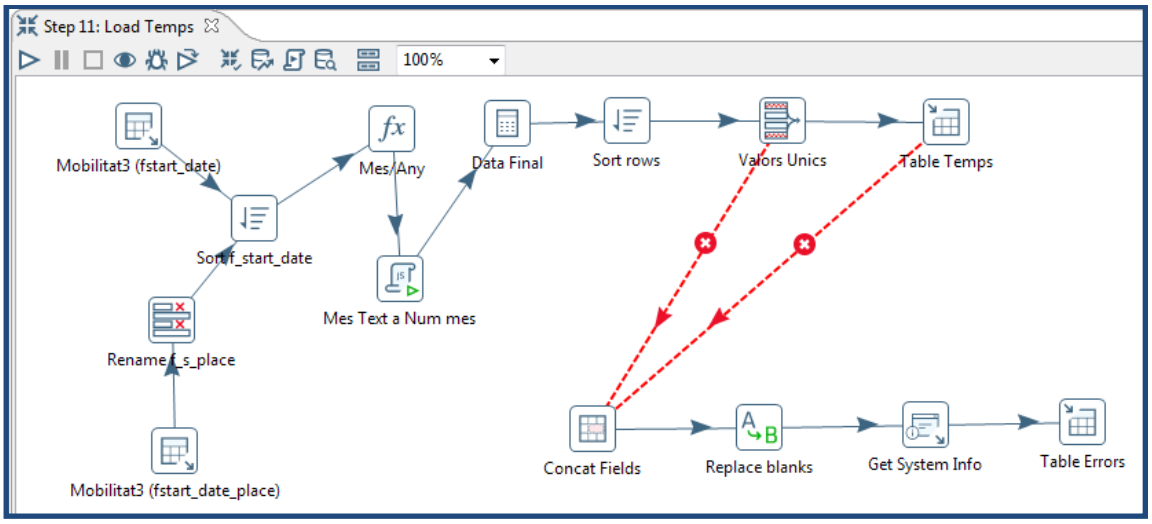


Figura 21. TRANS_INIT_Step11.ktr

Transformació 13 (TRANS_INIT_Step12.ktr): càrrega de dades mestres d'empreses a partir de la taula de dades transaccionals Mobilitat3, veure a la figura 22. Això ho hem de fer perquè no tenim dades mestres d'empreses. S'ha d'aplicar un *step* intermedi per eliminar la gran quantitat d'espais en blanc que conté aquest camp.

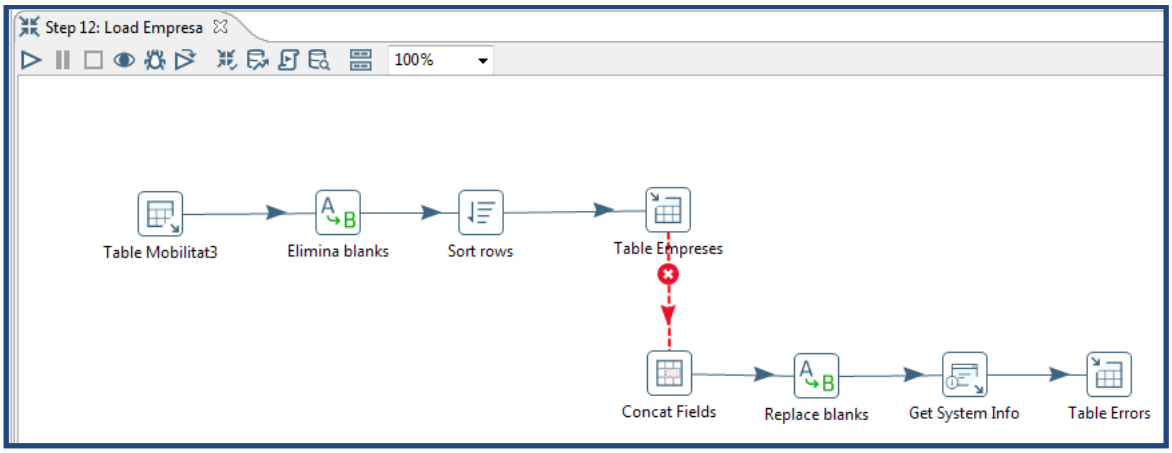


Figura 22. TRANS_INIT_Step12.ktr

Transformació 14 (TRANS_INIT_Step13.ktr): càrrega final a la taula de fets, veure a la figura 23. Un cop s'han omplert totes les taules de dimensions ens queda la taula de fets. A la taula de fets gravem principalment codis de taula *integer*, per tenir un rendiment òptim en l'accés al cub. Per tant la principal tasca d'aquesta transformació és utilitzar steps de *lookup* per buscar els codis únics i col·locar-los a la taula de fets.

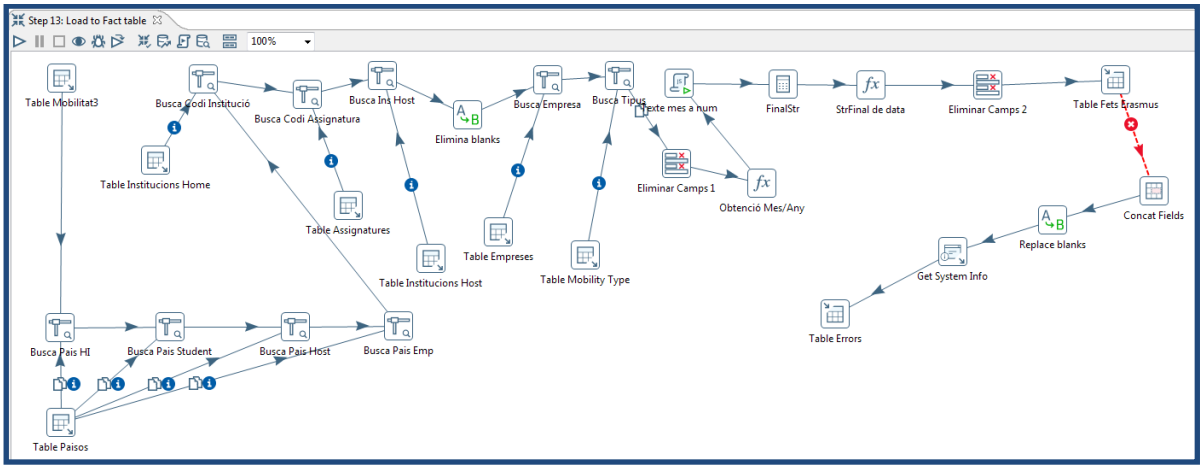


Figura 23. TRANS_INIT_Step13.ktr

Transformació 15 (TRANS_INIT_Step14.ktr): Finalment i per separar els registres nous dels ja afegits a la taula de fets des de Mobilitat3, actualitzem els registres de la taula Mobilitat3 marcant el camp fDelta, el codi es pot veure a la figura 24.

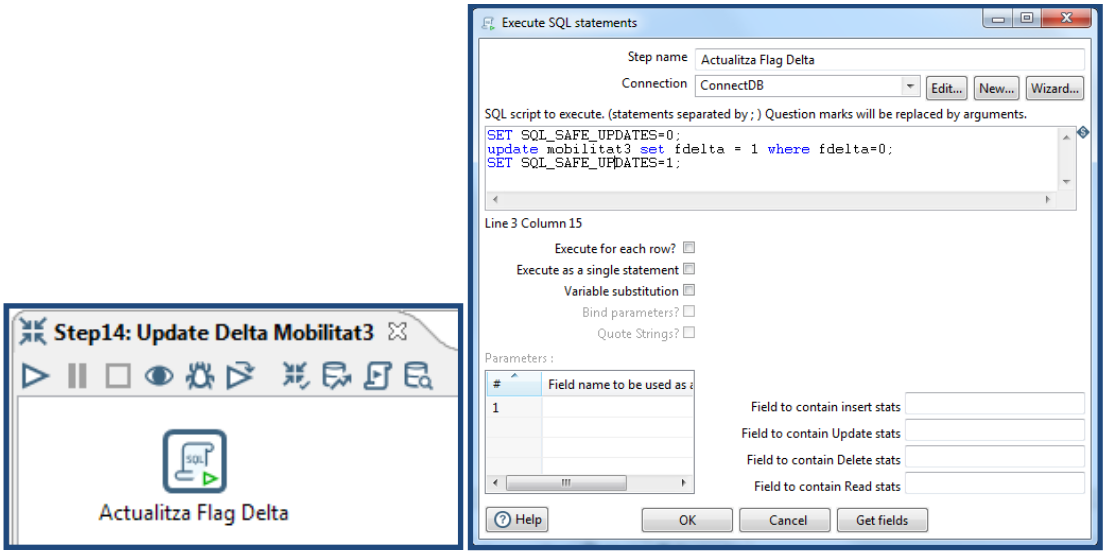


Figura 24. TRANS_INIT_Step14.ktr

La idea és que totes aquestes transformacions s'executin de manera automàtica i seqüencialment, inclús planificar el procés mensualment, anualment... Tot això ho fem amb els *jobs*. Per aquesta càrrega inicial tenim 3 que és mostrem les figures 25, 26 i 27:

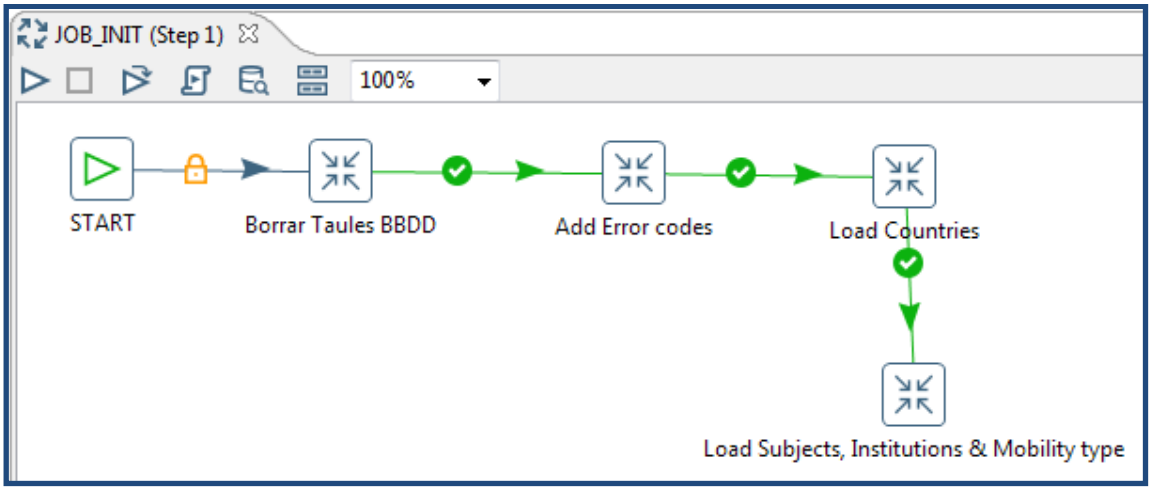


Figura 25. JOB_INIT(Step 1)

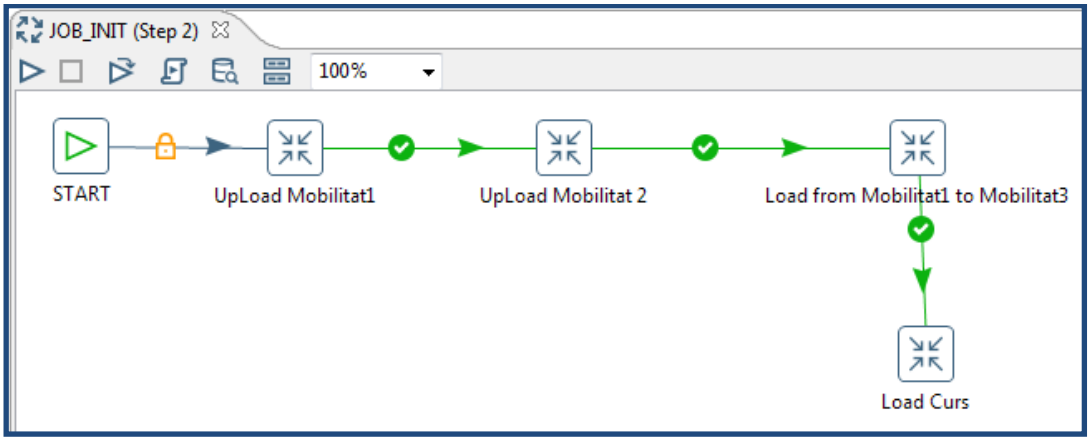


Figura 26. JOB_INIT(Step 2)

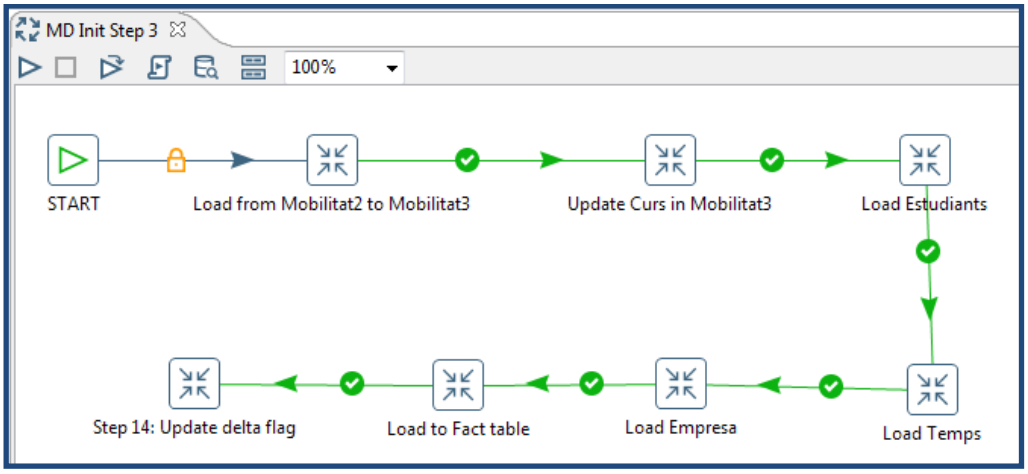


Figura 27. JOB_INIT(Step 3)

8. CÀRREGA DE DADES DELTA

Una vegada s'ha fet la càrrega de dades inicial el sistema ha d'estar preparat per futures càrregues de dades. Ara començarem mostrant el treball automàtic que fa aquesta tasca i després veurem en detall les diferents transformacions. En la càrrega delta tenim tant transformacions especials delta com aprofitament de transformacions *init* que serveixen també com a delta.

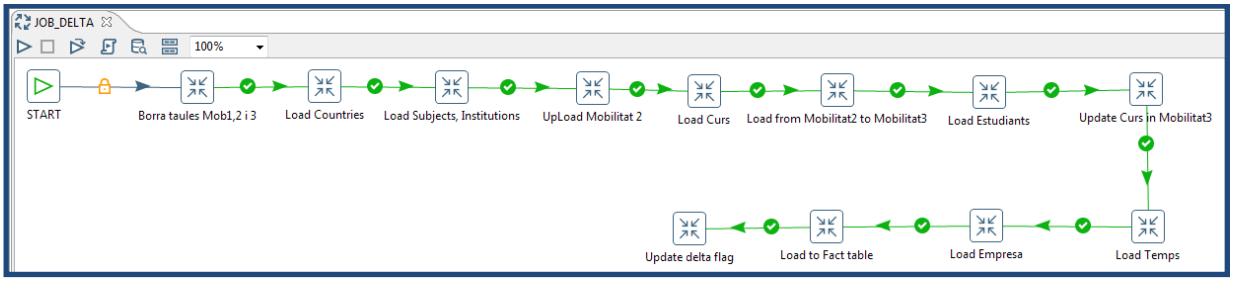


Figura 28. JOB_DELTA

Com podem veure a la figura 28 tenim 12 transformacions que veurem en detall a continuació.

Transformació 1 (TRANS_DELTA_Step1.ktr): Borrat de les taules temporals Mobilitat1 i Mobilitat2, el codi es pot veure a la següent figura 29.

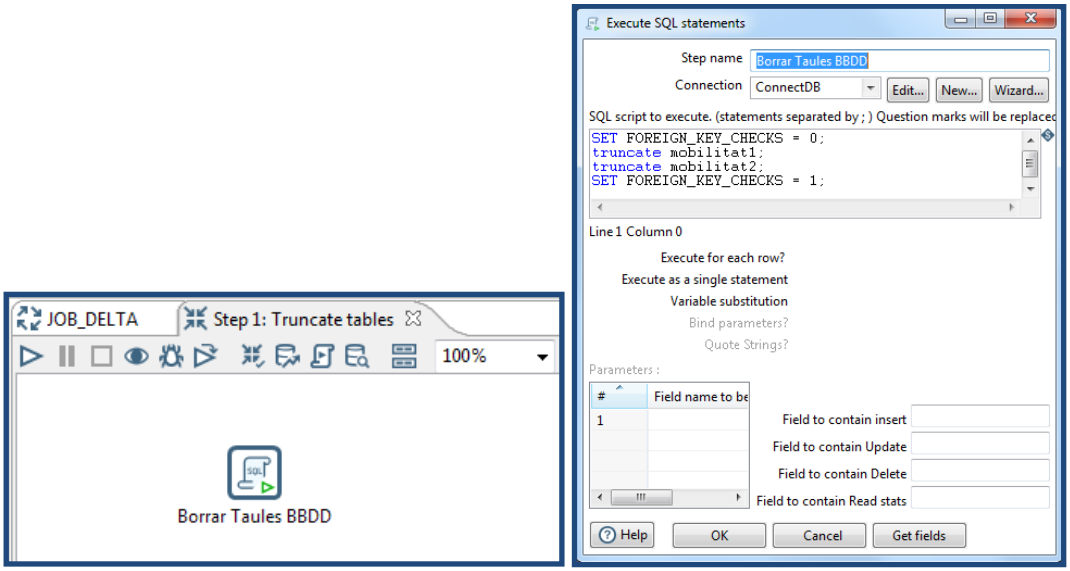


Figura 29. TRANS_DELTA_Step1.ktr

Transformació 2 (TRANS_DELTA_Step2.ktr): càrrega de dades mestres de països. L'única diferencia amb la corresponent transformació *Init* és que abans de gravar a la taula mira si el valor està ja guardat, per mitjà d'un "*stream lookup*", i sinó és així, el grava.

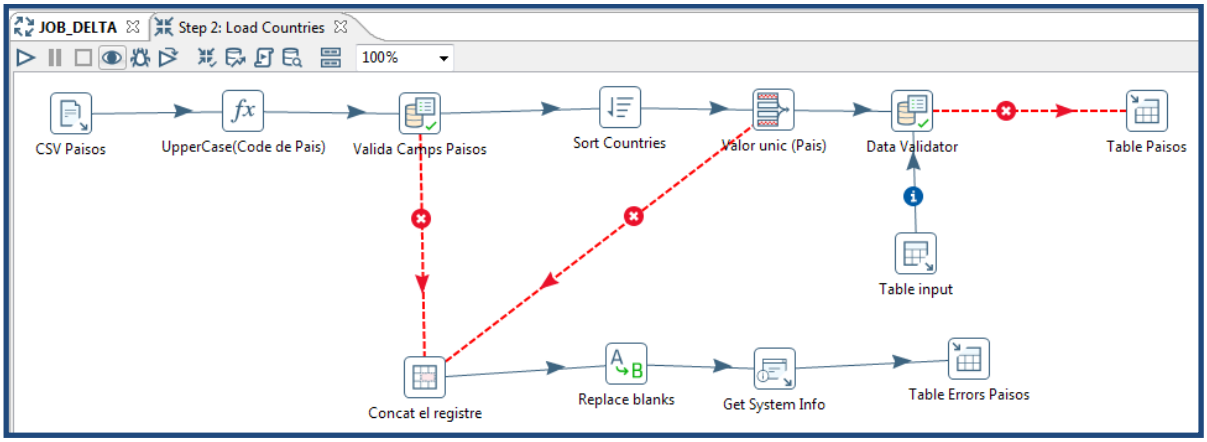


Figura 30. TRANS_DELTA_Step2.ktr

Transformació 3 (TRANS_DELTA_Step3.ktr): càrrega de dades mestres d'institucions i assignatures, veure figura 31. L'única diferencia amb la corresponent transformació *Init* és que abans de gravar a la taula mira si el valor està ja guardat i sinó és així, el grava.

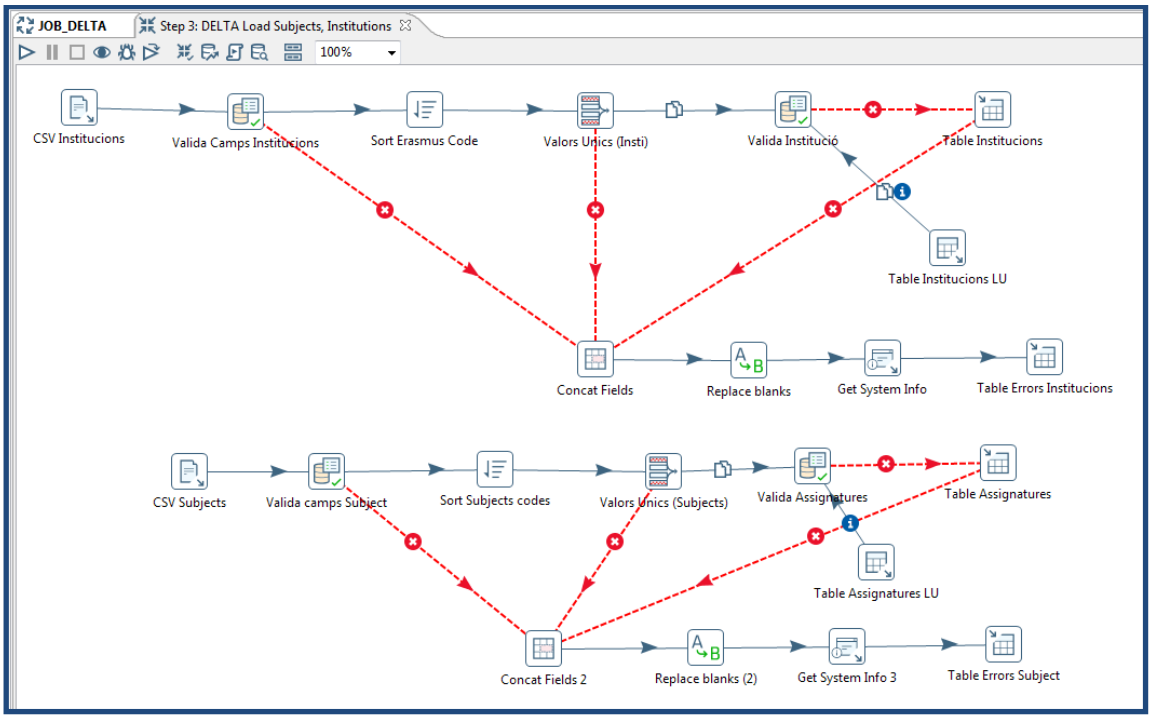


Figura 31. TRANS_DELTA_Step3.ktr

Transformació 4 (TRANS_DELTA_Step4.ktr): càrrega de dades transaccionals de curs vinents amb la diferència que en aquesta ocasió la captura del fitxer CSV es fa amb un nom genèric de fitxer, la resta és similar a la seva corresponent càrrega *Init*.

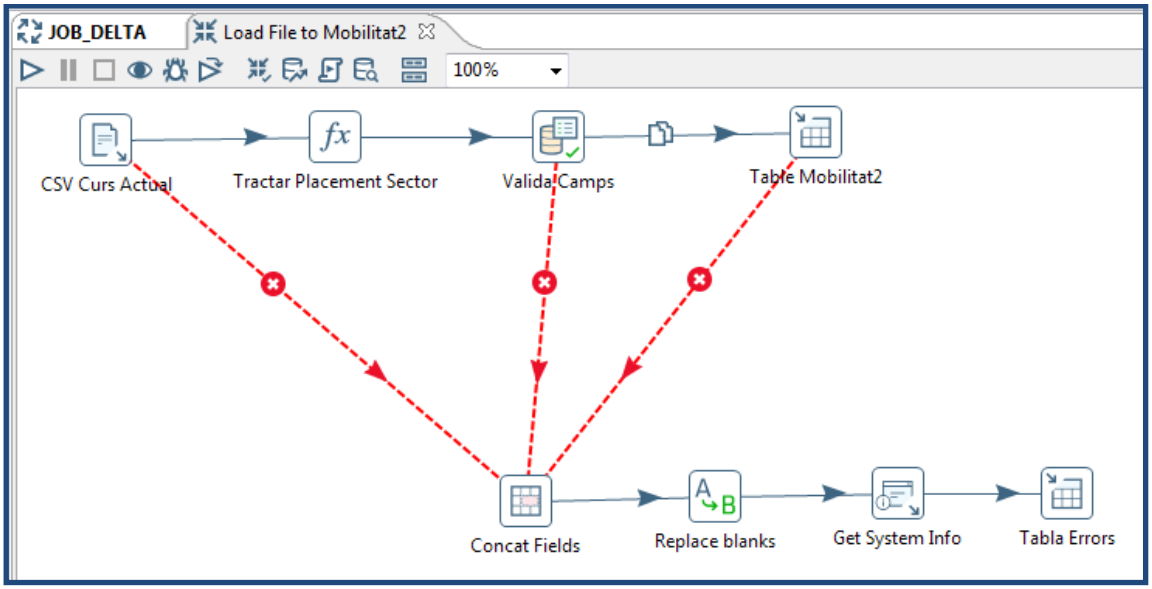


Figura 32. TRANS_DELTA_Step4.ktr

Transformació 5 (TRANS_INIT_Step7.ktr): càrrega de dades mestres curs, ja definida a l'etapa de càrrega inicial.

Transformació 6 (TRANS_INIT_Step9.ktr): càrrega de dades transaccionals de la taula de Mobilitat2 a 3, ja definida a l'etapa de càrrega inicial.

Transformació 7 (TRANS_DELTA_Step5.ktr): càrrega de dades mestres d'estudiants, però en aquest cas només llegim les dades de Mobilitat3 amb el camp "fDelta" = 0, que indica registres nous, per no repetir dades.

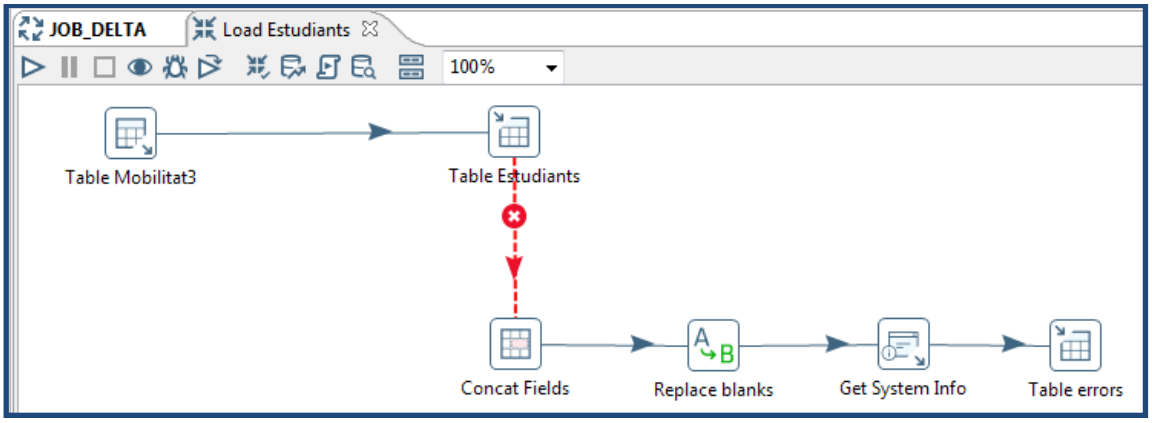


Figura 33. TRANS_DELTA_Step5.ktr

Transformació 8 (TRANS_INIT_Step8.ktr): actualització de camp "curs" a la taula Mobilitat3, ja definida a l'etapa de càrrega inicial.

Transformació 9 (TRANS_INIT_Step11.ktr): càrrega de dades mestres de Temps, ja definida a l'etapa de càrrega inicial.

Transformació 10 (TRANS_DELTA_Step6.ktr): càrrega de dades mestres d'empreses, avaluant registre per registre si ja està donat d'alta o no. Com en el cas dels estudiants només llegim les dades de la taula Mobilitat3 amb el camp "fDelta" = 0.

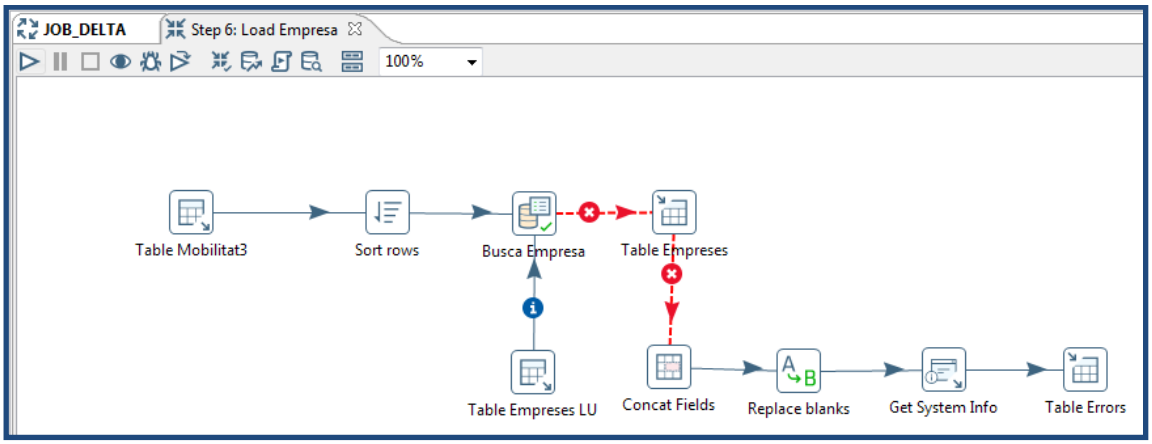


Figura 34. TRANS_DELTA_Step6.ktr

Transformació 11 (TRANS_INIT_Step13.ktr): càrrega de la taula de fets, ja definida a l'etapa de càrrega inicial.

Transformació 12 (TRANS_INIT_Step14.ktr): actualització de camp "fdelta" a la taula Mobilitat3, ja definida a l'etapa de càrrega inicial.

Nota: per fer les proves de dades deltes hem utilitzat una sèrie de fitxers, que podem veure en la figura 35, els quals han servit per validar el circuit o procés vist en aquest apartat. Per provar els fitxers s'han de col·locar on estan els de la càrrega inicial

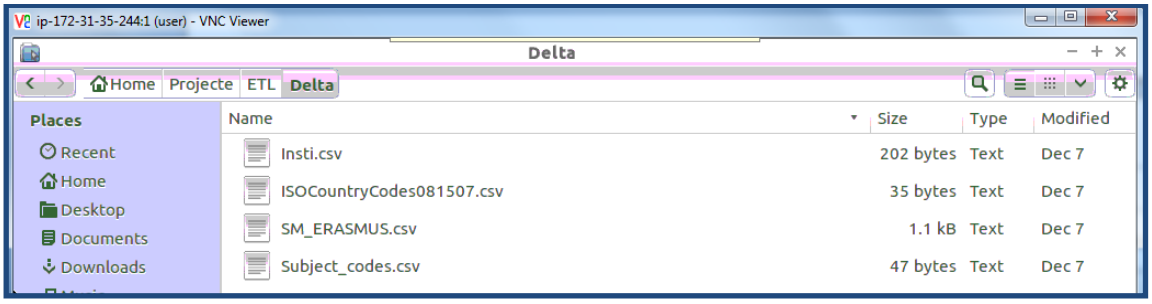


Figura 35. Fitxers DELTA

9. GESTIÓ D'ERRORS

Com s'ha anat veient en les transformacions anteriors, en cada una d'elles hi ha un control o gestió d'errors que a continuació descrivim.

En cada *step* susceptible d'error s'ha col·locat un *hop* que gestiona els errors, quan es detecta l'error s'envia el registre a un tractament posterior on concatenem tots els camps del registre, per poder estudiar l'error amb posterioritat, eliminem els espais en blanc per ocupar menys espai, llegim el dia i l'hora del sistema més el codi de transformació, per identificar amb més facilitat els errors. En la figura 36 podem veure un exemple.

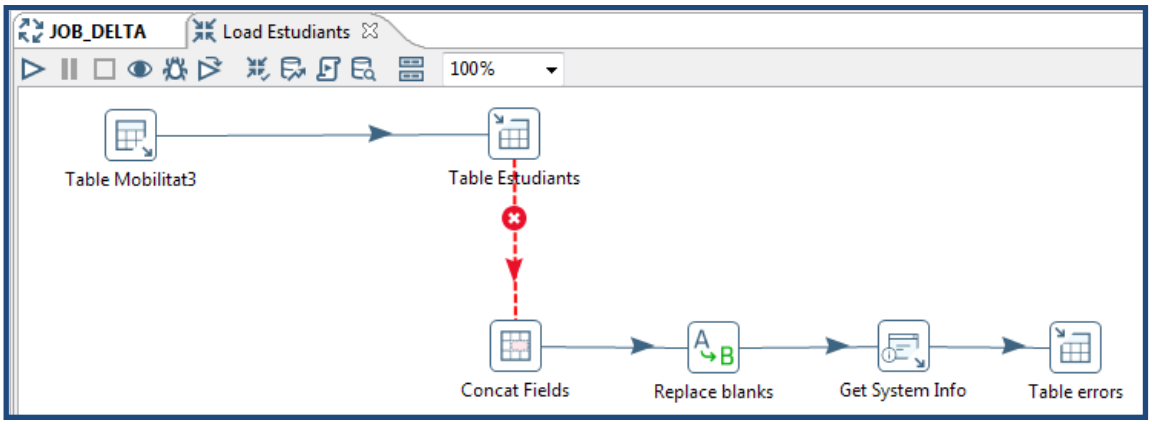


Figura 36. Gestió d'errors

El primer que fem és configurar el *hop* de gestió d'errors, marcant el destí de l'error, agafant el codi d'error i la descripció, com es veu a la figura 37:

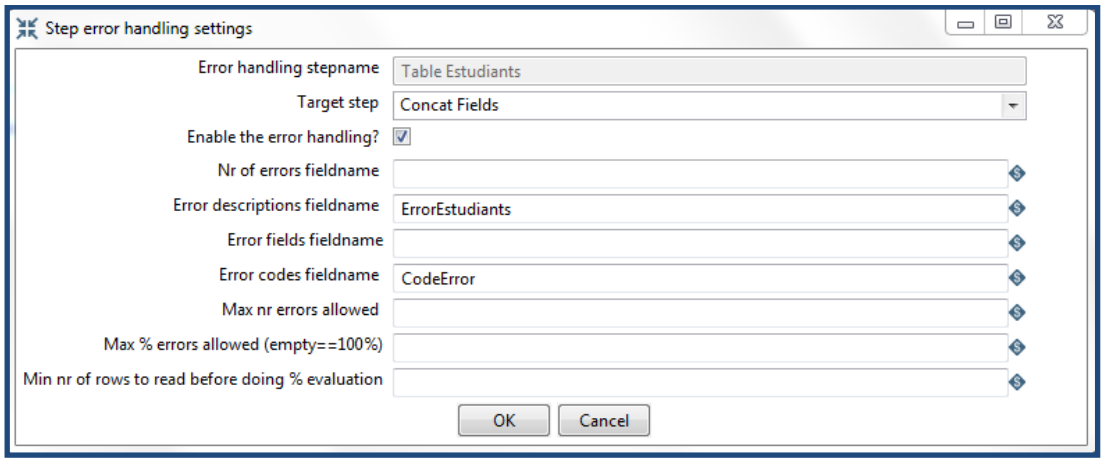


Figura 37. Gestió d'errors en *Hop*

Després concatenem tots els camps que han generat l'error, indicant el separador, i el camp on volem enviar la informació, en aquest cas "*row*", que apareix a l'apartat *Target field name* de la figura 38:

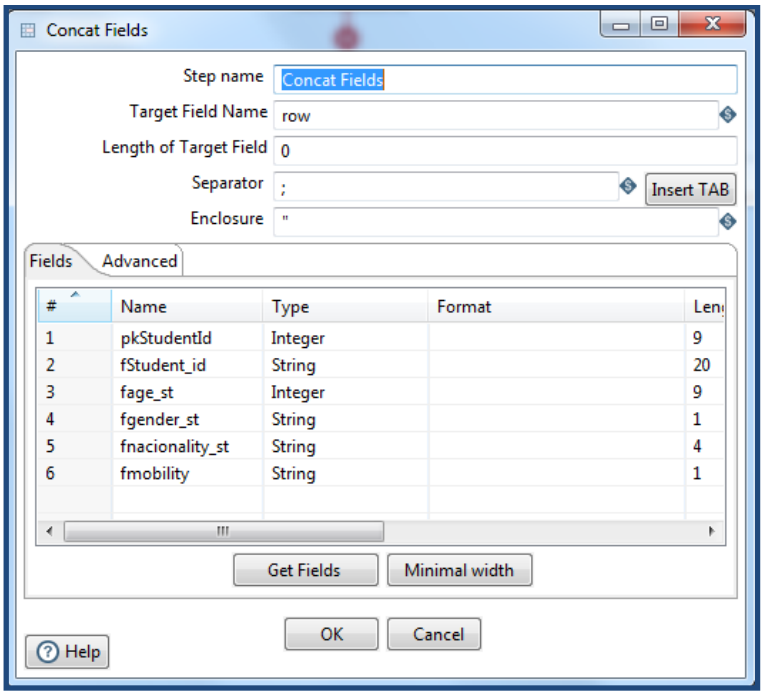


Figura 38. Concatenar camps

Eliminem els espais en blanc.

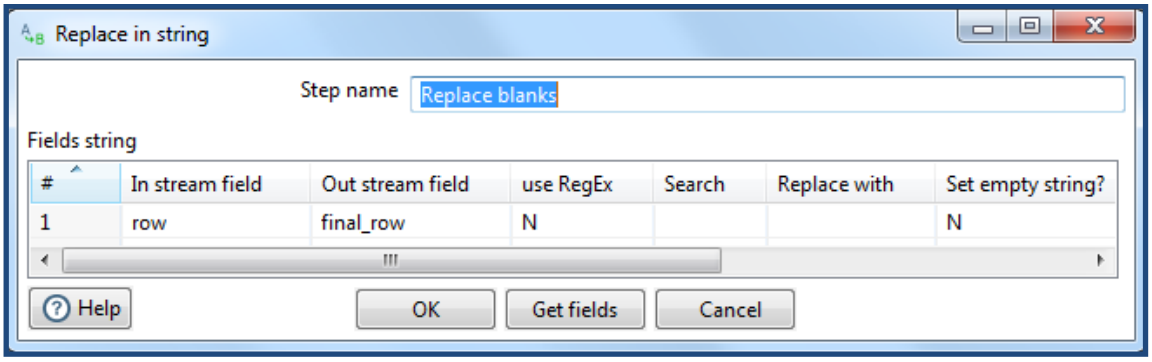


Figura 39. Eliminar blancs

A continuació agafem el codi de transformació i el dia i l'hora:

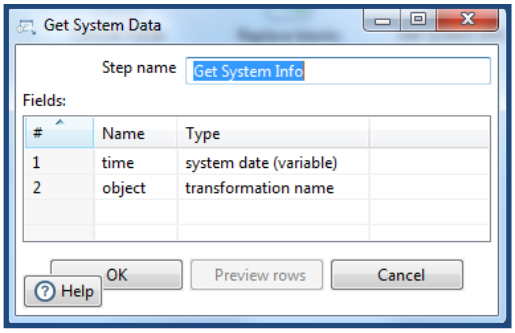


Figura 40. Informació de sistema

I finalment ho guardem tot en una taula que s'ha anomenat Errors:

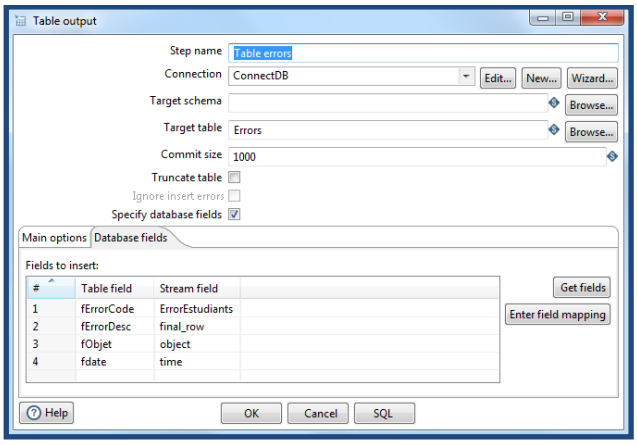


Figura 41. Taula d'errors

El propòsit final és tenir els errors en una taula, la qual es pugui consultar i inclús, en una fase futura evidentment, realitzar informes de gestió d'errors. La informació a consultar seria com la que es veu a la figura 42.

idErrors	fErrorCode	fErrorDesc	fObjet	fdate
1	Error Long Code Subject	Full;SubjectAreaDescription	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
2	Duplicate Subject code	142;Educationscience	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
3	Duplicate Subject code	144;Trainingforteachersatbasiclevels	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
4	Duplicate Subject code	144;Trainingforteachersatbasiclevels	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
5	Duplicate Subject code	212;Musicandperformingarts	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
6	Duplicate Subject code	214;Design	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
7	Duplicate Subject code	222;Foreignlanguages	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
8	Duplicate Subject code	222;Foreignlanguages	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
9	Duplicate Subject code	222;Foreignlanguages	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
10	Duplicate Subject code	222;Foreignlanguages	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
11	Duplicate Subject code	223;Mothertongue	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
12	Duplicate Subject code	223;Mothertongue	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
13	Duplicate Subject code	223;Mothertongue	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
14	Duplicate Subject code	223;Mothertongue	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
15	Duplicate Subject code	225;Historyandarchaeology	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
16	Duplicate Subject code	225;Historyandarchaeology	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
17	Duplicate Subject code	312;Sociologyandculturalstudies	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
18	Duplicate Subject code	312;Sociologyandculturalstudies	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
19	Duplicate Subject code	313;Politicalscienceandcivics	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39
20	Duplicate Subject code	321;Journalismandreporting	Step 3: Load Subjects, Institutions & Mobility type	2015-12-12 17:45:39

Figura 42. Exemple taula d'errors

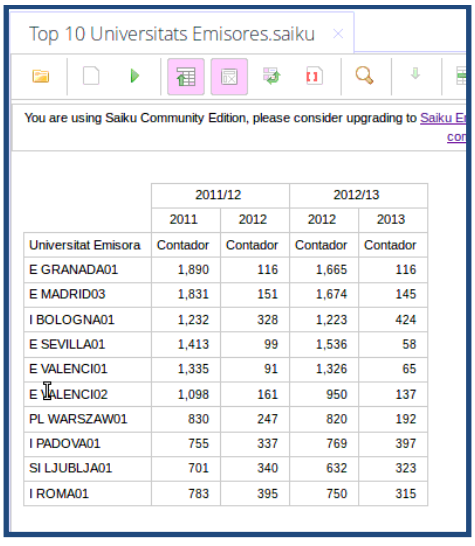
10. INFORMES

Els informes s'han realitzat amb *Saiku* de la solució *Pentaho*. Aquest és un entorn de desenvolupament d'informes basat en Web. És bastant *user-friendly* però amb algunes limitacions en la seva versió de *Community Edition*.

A continuació veurem un per un els informes realitzats, en l'opció de visualització de dades i l'opció gràfica.

10.1. Top 10 Universitats emissores

Informe on mostrem a les files, les universitats que més estudiants envien a Europa comparant l'any 2011/12 i 2012/13, a les columnes. Només mostrem les 10 universitats que més envien (Top 10), com es veu a la figura 43



	2011/12		2012/13	
	2011	2012	2012	2013
Universitat Emisora	Contador	Contador	Contador	Contador
E GRANADA01	1,890	116	1,665	116
E MADRID03	1,831	151	1,674	145
I BOLOGNA01	1,232	328	1,223	424
E SEVILLA01	1,413	99	1,536	58
E VALENCI01	1,335	91	1,326	65
E VALENCI02	1,098	161	950	137
PL WARSZAW01	830	247	820	192
I PADOVA01	755	337	769	397
SI LJUBLJA01	701	340	632	323
I ROMA01	783	395	750	315

Figura 43. Top 10 Universitats emissores

També podem veure l'informe en format gràfic, figura 44, a part de veure les dades. En aquest cas veiem un gràfic de barres en què cada color representa un curs diferent. A primer cop d'ull veiem que les universitats Espanyoles són de les que més estudiants envien liderada per la universitat de Granada.

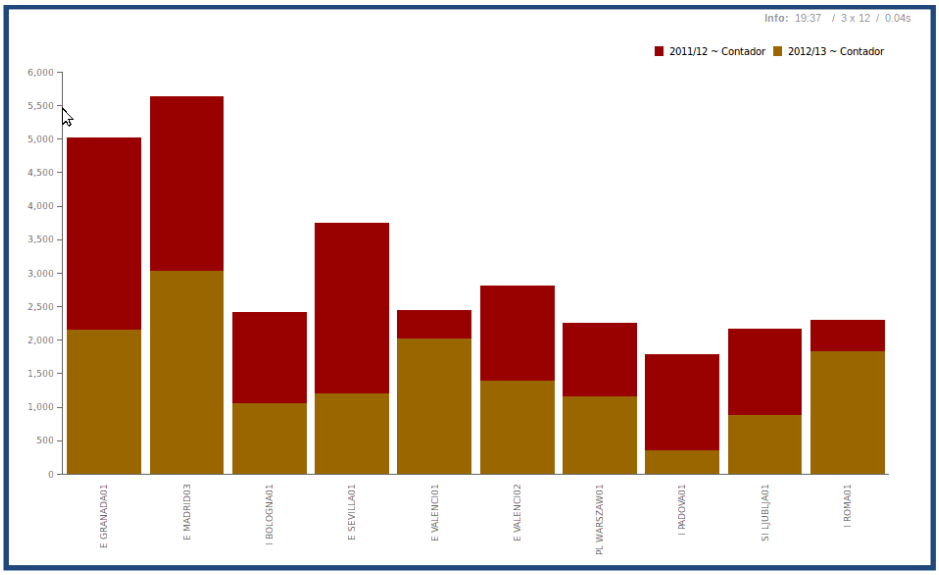


Figura 44. Gràfica Top 10 Uni. emissores

En aquest informe, com els que veurem a continuació excloem els valors incorrectes, és a dir, els valor que en el procés de càrrega de dades no tenen correctament les dades mestres. En aquest cas en concret, no volem veure les institucions amb codi "*XXXXXXXXXXXXXXX Unknown*".

10.2. Top 10 Universitats Receptores

Com en l'anterior informe fem una llista d'universitats, en aquest cas receptores. S'ordenen descendentment i es mostren les 10 primeres. A part, a les columnes afegim el curs i l'any a tall de comparació.

El que podem veure, tant a la figura 45 com 46 és de nou el predomini de les universitats espanyoles. Veient tant l'informe anterior com aquest sembla que les universitats més implicades o que més participen en el projecte Erasmus són clarament les espanyoles.

	2011/12		2012/13	
	2011	2012	2012	2013
Universitat Receptora	Contador	Contador	Contador	Contador
E GRANADA01	1,521	531	1,462	491
E VALENCI01	1,241	457	1,320	455
E SEVILLA01	1,359	410	1,301	389
E MADRID03	1,294	415	1,220	429
I BOLOGNA01	1,269	424	1,202	416
E VALENCI02	998	510	887	462
CZ PRAHA07	832	305	974	337
I ROMA01	880	227	874	259
E BARCELO01	740	365	724	371
E SALAMAN02	852	258	763	273

Figura 45. Top 10 Uni. receptores

També en aquest informe queden excloses les universitats que no tenen dades mestres.

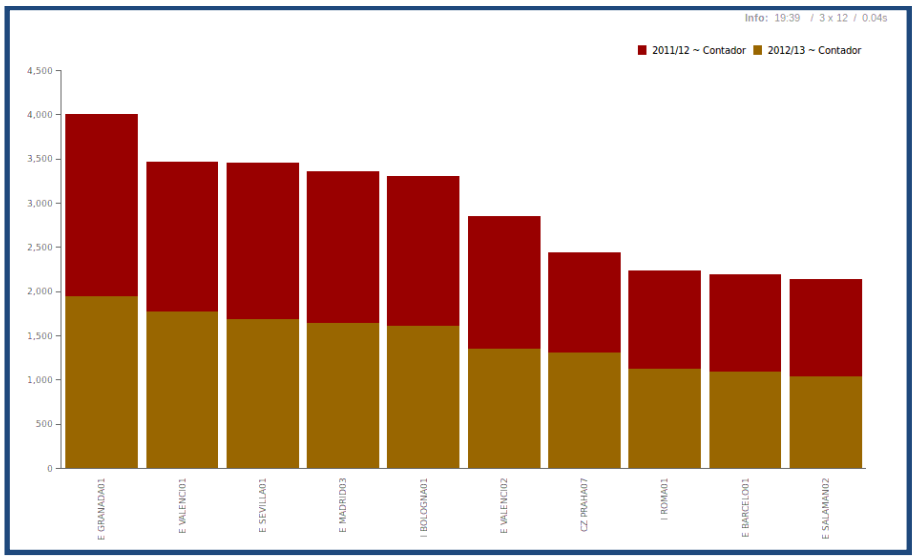


Figura 46. Gràfica Top 10 Uni. receptores

10.3. Estudiants per nacionalitat (%)

En aquest informe veiem per una banda totes les nacionalitats dels estudiants, i per una altra banda el pes que té cada país en forma de percentatge. Aquest percentatge s'ha hagut de calcular en script MDX directament en una fórmula calculada "Percet. Est(%)".

Info: 19:41 / 35 x 5 / 0.19s

	ES	DE	FR	IT	PL	TR	UK	NL	BE	PT	CZ	FI	RO	HU	AT
	Spain	Germany	France	Italy	Poland	Turkey	United Kingdom	Netherlands	Belgium	Portugal	Czech Republic	Finland	Romania	Hungary	Austria
Curs	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)
2011/12	15.25%	13.72%	12.93%	9.23%	6.15%	4.79%	4.40%	2.92%	2.70%	2.62%	2.50%	1.96%	1.91%	1.82%	1.78%
2012/13	14.27%	13.54%	12.87%	9.38%	6.03%	5.48%	4.43%	3.01%	2.78%	2.69%	2.43%	1.93%	1.97%	1.73%	1.73%

Figura 47. Estudiant per nacionalitat I (%)

S'ha cregut convenient afegir el Curs a mode comparatiu per veure les diferències entre els dos.

Info: 19:41 / 35 x 5 / 0.19s

GR	LT	SK	SE	IE	DK	BG	CH	LV	NO	SI	EE	HR	LU	CY	IS
Greece	Lithuania	Slovak Republic	Sweden	Ireland	Denmark	Bulgaria	Switzerland	Latvia	Norway	Slovenia	Estonia	Croatia	Luxembourg	Cyprus	Iceland
Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)	Percent. Est (%)
1.46%	1.48%	1.40%	1.30%	1.05%	0.98%	0.96%	0.94%	0.92%	0.69%	0.68%	0.44%	0.38%	0.17%	0.13%	0.12%
1.62%	1.43%	1.48%	1.33%	1.01%	1.01%	0.91%	0.89%	0.74%	0.66%	0.67%	0.42%	0.46%	0.16%	0.15%	0.11%

Figura 48. Estudiant per nacionalitat II (%)

En el gràfic de la figura 49 podem veure que és Espanya qui està al capdavant dels països en percentatges d'estudiants, la qual cosa veiem coherent, ja que Espanya és el país que més estudiants envia i rep al programa Erasmus, com hem vist als informes anteriors.

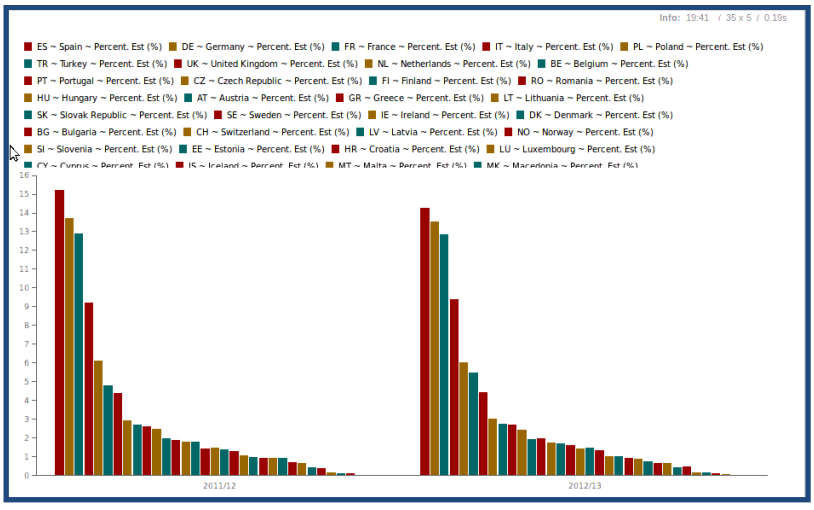


Figura 49. Gràfica Estudiant per nacionalitat (%)

En aquest informe s'ha exclòs el país "*XXXX Unknown*".

10.4. Estudiants per àrea de coneixement (%)

Informe igual que l'anterior però en lloc de fer el percentatge per nacionalitat ho fem per àrea de coneixement, és a dir assignatura, segon camp a la figura 50. A més es compara el percentatge del Curs 2011/12 amb el Curs 2012/13.

		2011/12	2012/13
Codi Assig.	Assignatura	Perce. Assg.(%)	Perce. Assg.(%)
222	Foreign languages	9.63%	8.31%
34	Business and administration	8.66%	7.83%
340	Business and administration (broad programmes)	7.29%	7.51%
52	Engineering and engineering trades	3.88%	4.03%
38	Law	4.10%	3.77%
314	Economics	3.76%	3.25%
22	Humanities	2.35%	3.93%
313	Political science and civics	3.22%	2.91%
721	Medicine	2.61%	2.26%
581	Architecture and town planning	2.73%	2.03%
31	Social and behavioural science	1.76%	2.20%
345	Management and administration	1.83%	1.94%
481	Computer science	1.80%	1.47%
58	Architecture and building	1.20%	2.01%
223	Mother tongue	1.62%	1.55%
421	Biology and biochemistry	1.66%	1.49%
521	Mechanics and metal work	1.56%	1.42%
380	Law	1.37%	1.59%
214	Design	1.47%	1.46%
32	Journalism and information	1.47%	1.36%
311	Psychology	1.46%	1.35%
812	Travel, tourism and leisure	1.25%	1.34%
582	Building and civil engineering	1.20%	1.18%
225	History and archaeology	1.25%	1.10%
723	Nursing and caring	1.14%	1.14%

Figura 50. Estudiants per àrea de coneixement (%)

Al gràfic de la figura 51 mostrem les dades que hem vist prèviament a la taula i podem veure en un primer cop d'ull que l'assignatura llengües estrangeres i les assignatures de Business (suposem direcció d'empresa) estan molt destacades en vers a totes les altres assignatures.

En aquest informe hem exclòs el codi d'assignatura "*XXXX Unknown*".

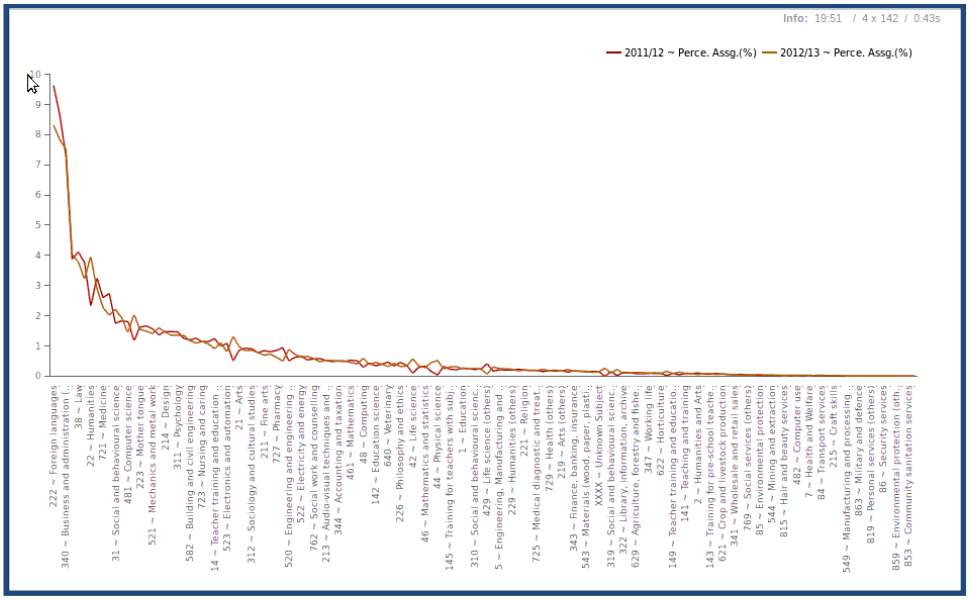


Figura 51. Gràfica estudiants per àrea de coneixement (%)

10.5. Evolució comparativa d'estudiants per Curs

Informació d'estudiants comparativa, en aquest cas comparem les 3 mesures que tenim al cub com és; nombre d'estudiants (Contador), beques (*Grant*), mitjana d'edat (Edat). Mirem l'evolució de les dades en els dos cursos que tenim actualment i hem afegit també el tipus de mobilitat a la posició de les columnes.

D'aquesta manera podem veure a la figura 52 que la gran majoria d'estudiants cursen en el tipus de mobilitat "S" de només mobilitat entre universitats. També podem veure amb les dades que tenim que el nombre d'estudiants d'un curs a un altre augmenta en uns 7000 estudiants entre tots els països.

	C			P			S		
	Study mobility combined with a placement			Placement mobility			Study mobility only		
Curs	Contador	Grant	Edat	Contador	Grant	Edat	Contador	Grant	Edat
2011/12	438	837,581.36	22.826	48,083	0	22.786	204,306	299,117,150.73	22.46
2012/13	476	1,031,762	22.527	55,481	0	22.877	211,221	330,729,579.146	22.418

Figura 52. Evolució comparativa d'estudiants per curs

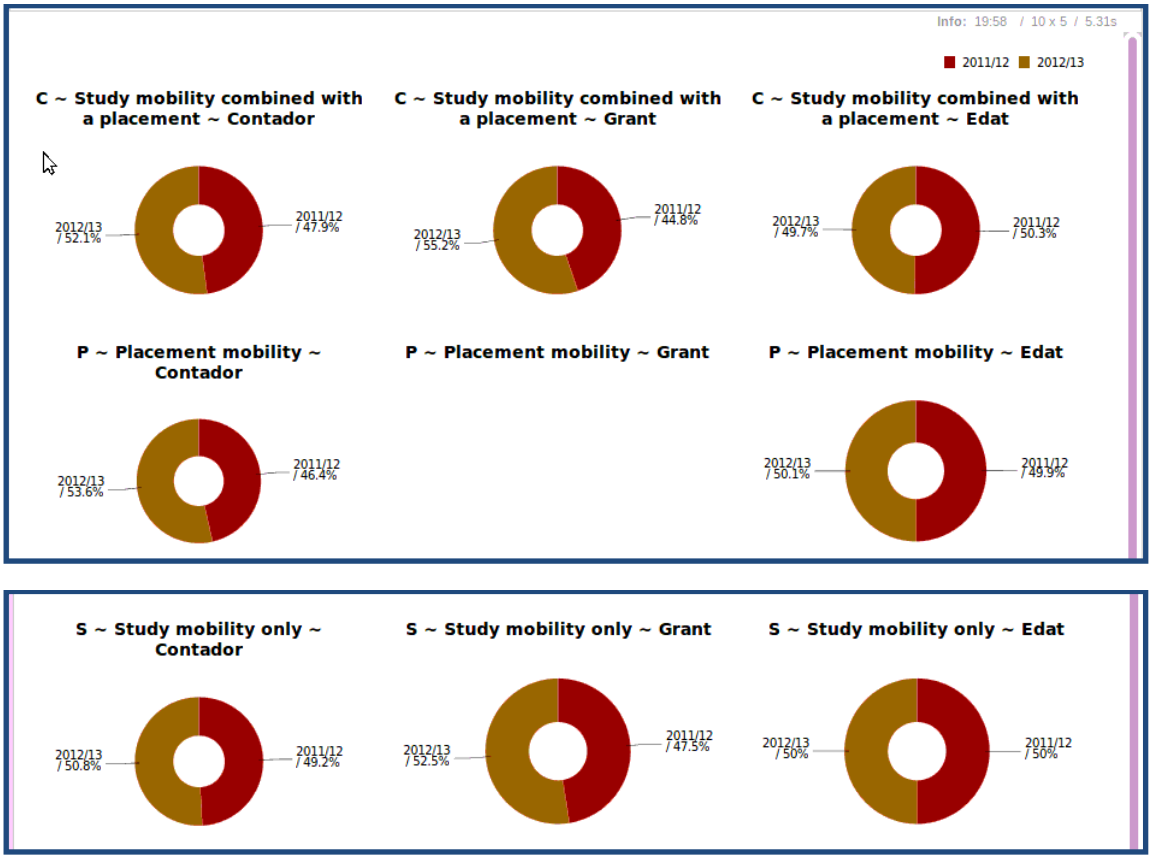


Figura 53. Gràfica evolució comparativa d'estudiants per curs

En el gràfic de la figura 53 podem veure que tant el nombre d'alumnes com les beques han augmentat al curs 2012/13 en tots els tipus de mobilitat. L'edat més o menys es manté igual, cosa lògica d'altra banda.

10.6. Edat mitjana d'estudiants per nacionalitat emissora

Llista de països que participen en el programa d'Erasmus per mostrar l'edat mitjana per nacionalitat. Hem ordenat les dades en ordre ascendent d'edat. Podem veure a la figura 54 que l'edat mitjana no canvia gaire d'un curs a un altre.

		2011/12	2012/13
Pais Home	Desc. Pais Home	Edat	Edat
CY	Cyprus	20.93	21.046
UK	United Kingdom	21.319	21.218
MT	Malta	21.383	21.197
FR	France	21.318	21.295
LT	Lithuania	21.527	21.547
IE	Ireland	21.617	21.838
NL	Netherlands	21.991	22.004
LU	Luxembourg	22.049	22.03
TR	Turkey	22.075	22.087
PT	Portugal	22.112	22.14
GR	Greece	22.206	22.363
RO	Romania	22.345	22.516
PL	Poland	22.505	22.544
ES	Spain	22.69	22.6
LV	Latvia	22.444	22.941
BG	Bulgaria	22.55	22.786
SK	Slovak Republic	22.722	22.743
HU	Hungary	22.786	22.828
EE	Estonia	22.923	23.143
SI	Slovenia	22.974	23.187
IT	Italy	23.083	23.131
HR	Croatia	23.386	23.158
CZ	Czech Republic	23.293	23.277
AT	Austria	23.367	23.429
DE	Germany	23.43	23.386
CH	Switzerland	23.436	23.473
FI	Finland	23.657	23.703
NO	Norway	23.698	23.714
SE	Sweden	23.969	23.856
DK	Denmark	23.86	23.964
IS	Iceland	25.556	25.329
LI	Liechtenstein	25	26.154

Figura 54. Edat mitjana per nacionalitat emissora

També es pot observar que el rang d'edats a tots els països va des dels 20 anys de Xipre fins als 26 de Liechtenstein. A la figura 55 podem veure les dades representades gràficament.

En aquest informe hem exclòs les dades del país "*XXXX Unknown*".

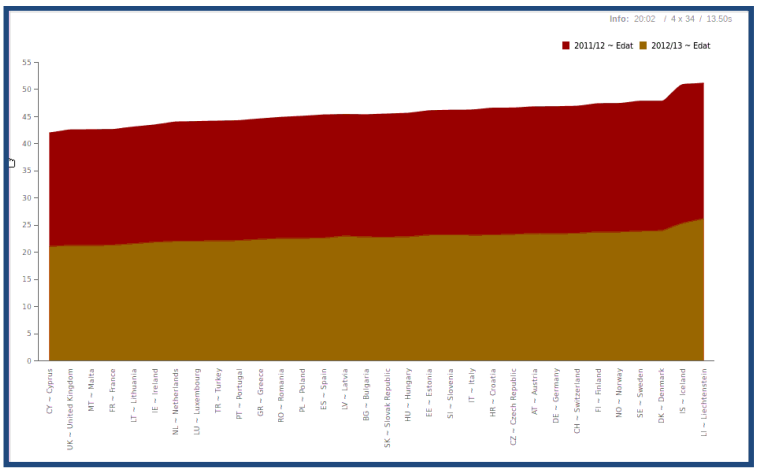


Figura 55. Gràfica Edat mitjana per nacionalitat emissora

10.7. Edat mitjana d'estudiants per nacionalitat receptora

Mateix cas que el cas anterior però basat en països receptors. També excloem el país "*XXXX Unknown*". Podem veure les dades a la figura 56 i el gràfic a la figura 57.

		2011/12	2012/13
Pais Receptor	Desc. Pais Receptor	Edat	Edat
IE	Ireland	22.11	22.142
CY	Cyprus	22.199	22.257
UK	United Kingdom	22.274	22.204
ES	Spain	22.319	22.285
LT	Lithuania	22.396	22.36
FR	France	22.426	22.34
NL	Netherlands	22.422	22.355
IT	Italy	22.47	22.42
PL	Poland	22.53	22.381
DK	Denmark	22.47	22.433
HU	Hungary	22.475	22.438
DE	Germany	22.478	22.45
CZ	Czech Republic	22.498	22.46
FI	Finland	22.503	22.484
LV	Latvia	22.514	22.481
BG	Bulgaria	22.739	22.397
SK	Slovak Republic	22.602	22.514
HR	Croatia	22.574	22.562
SI	Slovenia	22.589	22.554
SE	Sweden	22.585	22.571
GR	Greece	22.565	22.603
MT	Malta	22.699	22.51
NO	Norway	22.569	22.652
PT	Portugal	22.724	22.667
AT	Austria	22.721	22.701
RO	Romania	22.748	22.741
CH	Switzerland	22.788	22.752
EE	Estonia	22.761	22.778
TR	Turkey	22.778	22.776
LU	Luxembourg	22.244	23.43
IS	Iceland	23.002	22.844
LI	Liechtenstein	22.954	23.342

Figura 56. Edat mitjana d'estudiants per nacionalitat receptora

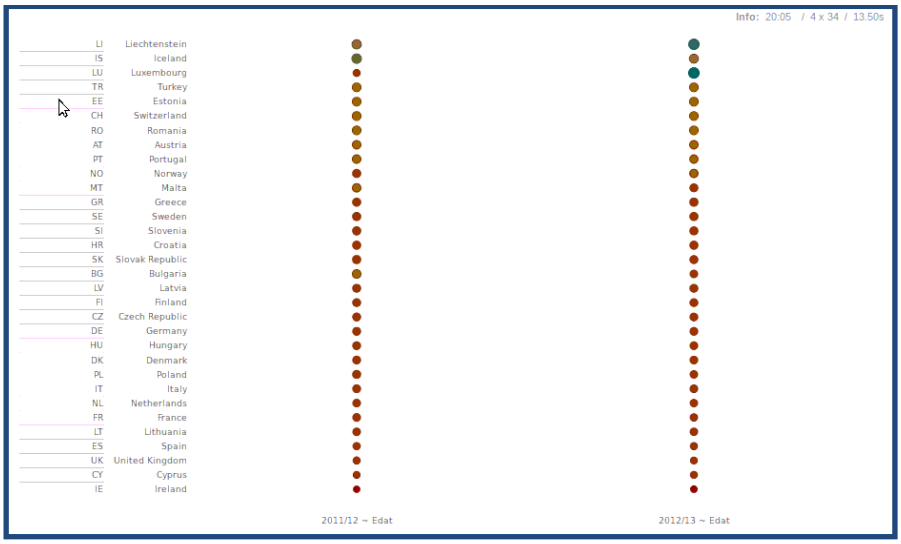


Figura 57. Gràfica Edat mitjana d'estudiants per nacionalitat receptora

10.8. Mitjana de beques per nacionalitat emissora

El que analitzem en aquest cas és un terme mitjà de beques en moneda per nacionalitat emissora, és a dir quin és el país que aporta més beques com a mitjana. Per portar a terme aquest requeriment s'ha creat un camp calculat al cub de tipus "*Average*" i el mostrem per nacionalitat emissora descendentment, com és por veure a la figura 58.

		2011/12	2012/13
Pais Home	Desc. Pais Home	Avg. Grant	Avg. Grant
IS	Iceland	2,391.628	2,743.494
CY	Cyprus	2,792.858	2,355.892
LI	Liechtenstein	2,306.489	2,707.385
BG	Bulgaria	2,422.185	2,498.023
TR	Turkey	2,054.958	1,986.01
EE	Estonia	1,967.249	2,027.336
RO	Romania	2,007.66	1,956.876
NO	Norway	1,675.937	2,132.24
GR	Greece	1,943.584	1,852.846
SE	Sweden	1,862.044	1,896.821
PL	Poland	1,841.698	1,824.47
UK	United Kingdom	1,739.393	1,701.85
SK	Slovak Republic	1,676.394	1,653.726
HR	Croatia	1,588.039	1,696.429
HU	Hungary	1,528.897	1,674.469
SI	Slovenia	1,524.152	1,529.753
LU	Luxembourg	1,489.171	1,425.991
MT	Malta	1,478.437	1,434.626
CH	Switzerland	1,611.541	1,121.09
PT	Portugal	1,228.869	1,225.183
LT	Lithuania	1,146.152	1,258.472
IT	Italy	1,176.392	1,210.968
IE	Ireland	1,112.974	1,217.778
CZ	Czech Republic	1,076.371	1,190.917
FI	Finland	994.275	1,085.82
LV	Latvia	1,050.477	962.682
DE	Germany	931.176	1,045.845
AT	Austria	874.717	973.592
FR	France	868.478	915.428
DK	Denmark	824.33	912.673
ES	Spain	763.95	872.274
NL	Netherlands	723.023	705.728

Figura 58. Mitjana beques per nacionalitat emissora

Per les dades vistes a les dades de la figura 58 i que es pot observar més ràpidament al gràfic de la figura 59 s'identifiquen dos límits. El límit superior o país que més aporta en beques que és Islàndia i el límit inferior o país que menys aporta que és Holanda.

Cal remarcar també que Espanya que és líder en enviament d'estudiants, curiosament és el segon per la cua en ajudes a estudiants, tot i que ha incrementat en un 14% la quantitat d'un curs a l'altre. En aquest informe també excloem el país "*XXXX Unknown*"

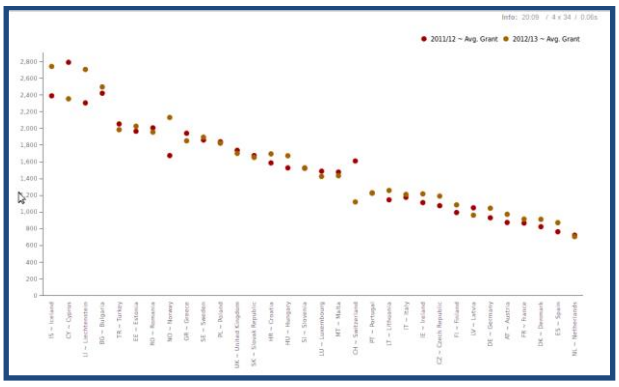


Figura 59. Gràfica mitjana beques per nacionalitat emissora

10.9. Mitjana de beques per nacionalitat receptora

Similar a l'informe anterior però en lloc d'analitzar nacionalitats emissores, aquí analitzem nacionalitats receptores. Les dades i gràfica es poden consultar a les figures 60 i 61. Com a l'informe per nacionalitat emissora, aquí també excloem el país "*XXXX Unknown*"

		2011/12	2012/13
Pais Receptor	Desc. Pais Receptor	Avg. Grant	Avg. Grant
LI	Liechtenstein	1,812.089	1,814.421
DE	Germany	1,721.836	1,819.961
FR	France	1,591.773	1,734.361
ES	Spain	1,567.11	1,652.991
DK	Denmark	1,531.764	1,663.097
AT	Austria	1,521.666	1,632.052
NL	Netherlands	1,461.924	1,575.805
IT	Italy	1,452.015	1,586.974
GR	Greece	1,426.151	1,622.364
PT	Portugal	1,457.278	1,559.272
CZ	Czech Republic	1,439.844	1,502.576
SI	Slovenia	1,383.811	1,537.564
CH	Switzerland	1,370.043	1,547.641
CY	Cyprus	1,337.636	1,541.316
HR	Croatia	1,307.105	1,504.044
MT	Malta	1,350.2	1,479.049
SK	Slovak Republic	1,373.061	1,457.938
SE	Sweden	1,364.691	1,464.908
HU	Hungary	1,401.128	1,414.955
PL	Poland	1,390.38	1,403.142
LU	Luxembourg	1,423.455	1,340.87
NO	Norway	1,301.387	1,446.558
LT	Lithuania	1,328.757	1,389.234
UK	United Kingdom	1,265.532	1,380.013
LV	Latvia	1,256.38	1,353.667
IS	Iceland	1,222.692	1,372.45
TR	Turkey	1,234.087	1,341.291
FI	Finland	1,233.713	1,343.178
EE	Estonia	1,262.534	1,312.697
BG	Bulgaria	1,255.439	1,306.547
RO	Romania	1,215.892	1,338.99
IE	Ireland	1,195.919	1,284.138

Figura 60. Mitjana de beques per nacionalitat receptora

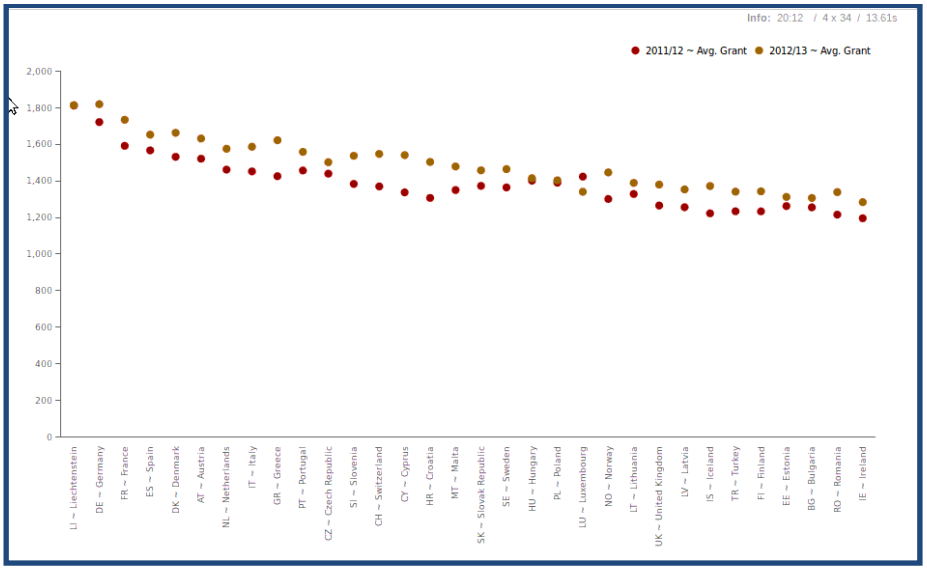


Figura 61. Gràfica mitjana de beques per nacionalitat receptora

11. SUPOSICIONS

La primera suposició i potser la més important és referent a l'estudiant. La primera idea era identificar els estudiants utilitzant el camp "*StudentId"*, però analitzant les dades hem vist que al Curs 2011/12 aquest camp estava en blanc. D'altra banda el Curs 2012/13 sí que tenia informat aquest camp però analitzant les dades massivament hem vist que les dades es repetien diverses vegades i amb edats diferents per tant podem dir que no era el mateix estudiant.

Al final s'ha decidit identificar cada registre de les dades transaccionals com un estudiant únic, tot i que les regles d'Erasmus fan menció que un alumne pot tenir més d'una línia, és el cas d'un estudiant que cursa més d'una assignatura, però es fa impossible identificar estudiants amb la qualitat de dades que tenim.

Després s'han introduït codis d'error a les taules, del tipus "*XXXX País Unknown*", d'aquesta manera no rebutjar els registres "defectuosos". Així aquests registres arriben al cub i en els informes es poden excloure o si es vol analitzar el perquè poder-los treballar.

S'ha intentat aplicar totes les regles de "*Student mobility_datadictionary.pdf*", és a dir, per exemple al camp "*Gender*" sexe en angles només s'accepten valors M o F...

Només s'ha fet l'excepció al camp "*short duration*" que en principi accepta els valors T o X, però la majoria de registres no complien aquesta regla i com no era un camp rellevant, s'ha decidit no aplicar-la, per no perdre tants registres.

A les dades mestres de països, s'han inserit 3 línies noves, que corresponen a Bèlgica, influenciada pels països Holanda, França i Alemanya. S'ha decidit fer això pel volum de dades transaccionals que tenen informat aquests països. Les línies són:

- BENL: Bèlgica(NL)
- BEFR: Bèlgica(FR)
- BEDE: Bèlgica(DE)

12. CONCLUSIONS

El procés d'elaborar un magatzem de dades és una tasca normalment complicada. Per aconseguir aquesta fita s'han passat per diferents fases.

Primerament vàrem fer una anàlisi preliminar i el pla de treball, on s'assentaven les bases de com seria el magatzem i s'establien les línies temporals del projecte. Tot això a partir d'un enunciat, veure annex, i unes fonts de dades proporcionades per Erasmus. La complicació en aquesta fase radicava en la poca informació proporcionada a l'enunciat. És aquí on es necessiten les suposicions, per a omplir els buits que deixava l'enunciat. Després s'ha vist en la planificació temporal del projecte el temps tan reduït que teníem per dur-lo a terme, per tant era important no perdre gaire temps, ni desviar-se dels temps marcats en el diagrama de Gannt.

A la segona part del projecte entràvem al detall de l'anàlisi i el disseny de la solució. Això significa intentar encertar amb la màxima precisió com serà la BBDD, el model multidimensional i saber quina és la informació aprofitable de les fonts de dades. Aquesta part del projecte també té la seva dificultat, ja que has de tenir una ment molt analítica i experiència, que també ajuda. Tot i així hem hagut de modificar certs aspectes de l'anàlisi inicial, en la fase d'implementació. Pocs projectes encerten el 100% de l'estudi teòric inicial amb el producte final que s'entrega.

Finalment s'ha procedit a la fase d'implementació, on els principals enemics que teníem eren el temps i la falta de coneixement de les eines utilitzades. Segurament amb més coneixement s'hauria entregat un millor producte, però al món real sovint et trobés en situacions similars i al final has d'aconseguir el teu objectiu que és fer un entregable robust i que cobreixi el màxim de requeriments del client. En aquest aspecte ha estat un bon exercici de cara al món laboral en el qual estem.

Després d'analitzar les diferents fases del TFC i veure el producte obtingut crec sincerament que s'han assolit tots els objectius del projecte, ja que s'ha construït un magatzem de dades amb el seu model multidimensional OLAP per explotar una sèrie d'informes i s'ha deixat tot preparat per rebre dades de curos vinents com es demanava a l'enunciat. També s'ha incorporat una taula d'errors on fer consultes i així poder gestionar adequadament el flux del procés ETL.

Tanmateix s'ha de dir que la part menys robusta són els informes, principalment per falta de temps, falta de coneixement i les poques característiques que ofereix l'eina *Saiku*.

Durant l'elaboració del projecte s'ha respectat el pla de treball en cada una de les seves fases. Malauradament el que no s'ha aproximat és el temps, ja que s'ha hagut de dedicar aproximadament un 15% més de temps de l'establert inicialment. Principalment l'increment de temps ha estat degut als canvis que s'han fet durant la implementació i que no estaven plantejats a les etapes inicials. Com s'ha comentat abans, alguns d'aquests canvis són falta de coneixement en l'ús de les eines.

En línies generals ha estat un bon producte i robust, amb parts millorables en el futur com són els informes i l'automatització de captura de fitxers, que ara és una tasca més manual.

13. GLOSSARI

BBDD: Acrònim de base de dades.

BI (Pentaho): És un conjunt de programes lliures per generar intel·ligència de negoci que inclouen eines integrades per generar informes, mineria de dades, processos ETL, etc...

CSV: *Comma-separated values*. Tipus d'arxiu editable des de qualsevol full de càlcul en el qual tots els registres que conté estan separats per comes (o punts i coma).

Dimensió: Punt de vista des del qual es poden analitzar les dades.

ETL*: Extract, Transform, Load.* S'utilitza per descriure els processos d'extracció, transformació i càrrega que permeten a les organitzacions moure dades des de múltiples fonts, formatjar-los, netejar-los i carregar-los en una altra base de dades.

Fets: Conjunt d'esdeveniments amb dades numèriques associades.

Font de dades: Conjunt de dades, en aquest cas en format de fitxer, que són el punt de partida del flux de dades d'un model.

Lubuntu: és una distribució oficial del projecte Ubuntu que té per lema "menys recursos i més eficiència energètica", usant el gestor d'escriptori LXDE. El nom de lubuntu és una combinació de LXDE i Ubuntu.

Mondrian: és una de les aplicacions més importants de la plataforma Pentaho BI. És un servidor OLAP *open source* que gestiona comunicació entre una aplicació OLAP (escrita en Java) i la base de dades amb les dades font.

Multidimensional: Les BBDD multidimensionals s'utilitzen principalment per crear aplicacions OLAP i poden veure's com a BBDD d'una sola taula, per cada dimensió tenen un camp (o columna), i un altre camp per cada mètrica o fet.

OLAP: és l'acrònim en anglès de processament analític en línia (*online analytical processing*). És una solució que subministra respostes ràpides de consultes a una base de dades.

Putty: és un client SSH, Telnet, rlogin i TCP raw amb llicència lliure.

Saiku: és una eina que permet dissenyar informes per explorar fonts de dades complexes usant una interfície gràfica amb la funcionalitat d'arrossegar i deixar, més conegut com *drag & drop*.

Script: És un arxiu d'ordres o de processament per lots que és interpretat per un intèrpret de comandes i s'utilitza per realitzar diverses tasques de forma seqüencial.

SQL: *Structured Query Language,* és un llenguatge declaratiu d'accés a BBDD relacionals que permet especificar diferents operacions en aquesta. Una de les seves característiques és l'ús de l'àlgebra i el càlcul relacional que permeten efectuar consultes amb la finalitat de recuperar, d'una manera senzilla, informació de BBDD, així com fer canvis.

SO: Acrònim de sistema operatiu.

Spoon: és el dissenyador gràfic de transformacions i treballs associat amb el sistema ETL Pentaho Data Integration, també conegut com a Kettle.

Schema-workbench: eina gràfica que permet la construcció dels esquemes de Mondrian i a més permet publicar-los al servidor de BI.

Star schema: és un model de dades que té una o més taules de fets que contenen les dades d'anàlisi, rodejada de taules de dimensió. Aquest aspecte d'una taula central rodejada de taules més petites és el que sembla una estrella, d'aquí el nom.

User-friendly: en aquest cas, és un programari de fàcil usabilitat, és a dir que no és gaire complicat de fer-ho servir.

XML: *Extensible Markup Language*. És un llenguatge de marques utilitzat per emmagatzemar dades en forma llegible. Aquest llenguatge dóna suport a bases de dades, sent així útil quan diverses aplicacions es volen comunicar entre si o integrar informació.

14. ANNEXOS

14.1. Annex 1. Enunciat projecte TFC

La Unió Europea ofereix estadístiques i fonts de dades en modalitat "open data" sobre diferents organismes i institucions europees. En el marc educatiu i dins de la Unió Europea en resulten d'especial interès les dades de mobilitat d'estudiants dins el programa Erasmus.

Aquestes dades que proporciona la Unió Europea permeten realitzar una anàlisi en profunditat sobre el moviment d'estudiants en base a diferents eixos d'anàlisi, com poden ser: nacionalitat (receptora i emissora), institució (receptora i emissora), edat, sexe, tipus de mobilitat, àrea de coneixement, etc.

L'objectiu d'aquest treball és integrar les fonts proporcionades per la Unió Europea amb l'objectiu de realitzar diferents tipus d'anàlisi, com poden ser:

Informes estàtics prefixats:

Top 10 d'Universitats més receptores, i més emissores d'estudiants Erasmus.
Distribució en % d'estudiants per nacionalitat.
Distribució en % d'estudiants per àrea de coneixement.
Evolució comparativa del nombre d'estudiants per curs.
Edat mitjana d'estudiants per nacionalitat receptora i emissor.
Quantitat mitjana de les beques per nacionalitat receptora i emissora.

Tots els indicadors anteriors han de poder ser analitzats comparant els diferents cursos (anys).

A més es podran afegir els informes que es considerin necessaris i que puguin interessar.

Informes lliures:

Les dades proporcionades permeten identificar una sèrie de dimensions d'anàlisi: edat, nacionalitat (receptora i emissora), sexe, curs i àrea de coneixement. Totes aquestes dades poden ser analitzades des de diverses dimensions. És per això que una anàlisi multidimensional amb eines OLAP pot ser molt útil per aquest tipus d'informes, ja que

permetrà afegir i desagregar per les dimensions d'anàlisi i estudiar el nombre d'estudiants des de totes aquestes dimensions.

15. BIBLIOGRAFIA

https://www.youtube.com/watch?v=kbgiwFNxsG4 **[Tutorial GanntProject]**

http://www.uoc.edu **[Portal UOC]**

http://lubuntu.es/ **[Sistema operatiu Ubuntu]**

https://www.mysql.com/ **[SGBD]**

http://www.pentaho.com/ **[Suit BI]**

http://community.pentaho.com/projects/data-integration/ **[Kettle, eina ETL]**

http://mondrian.pentaho.com/documentation/schema_workbench.pdf **[Manual modelatje]**

https://www.softcatala.org/corrector **[Corrector Ortogràfic]**

http://askubuntu.com/questions/407217/how-can-i-see-my-system-details-in-a-graphical-way-in-lubuntu **[Propietats Lubuntu]**

https://help.ubuntu.com/community/Lubuntu/Documentation **[Documentació Lubuntu]**

http://es.ccm.net/contents/311-comandos-de-linux **[Tutorial Linux]**

http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/ii-hefesto-metodologia-propia-para-la-construccion-un-data--13 **[Exemple relació m:n]**

https://www.youtube.com/watch?v=e0BUUlNuZQ8 **[Handling Many to Many Relationships Using Cubes in PDI and Pentaho Analyzer]**

http://forums.pentaho.com/showthread.php?148126-handling-many-to-many-relations-in-mondrian **[Handling many to many in mondorian]**

http://type-exit.org/adventures-with-open-source-bi/2010/08/data-validation-and-monitoring-with-pentaho-kettle/ **[Manual Validació dades amb Pentaho Spoon]**

http://mysql.rjweb.org/doc.php/datawarehouse **[Manual Data Warehouse]**

https://www.youtube.com/watch?v=baC8dQMyp9E **[Vídeo d'us schema-workbench]**

https://www.youtube.com/watch?v=AOUCN5HxmX4 **[Vídeo d'us schema-workbench]**

http://www.businessintelligence.info/serie-dwh/tablas-de-hecho-fact-tables.html **[Manual Taules de fets]**

http://www.businessintelligence.info/serie-dwh/claves-subrogadas.html **[Manual claus foranes]**

http://mondrian.pentaho.com/documentation/schema.php#Degenerate_dimensions **[Documentació Modrian Pentaho]**