Anàlisi de mobilitat d'estudiants Erasmus (Construcció i explotació d'un Data Warehouse)

Jordi Feliu Sobré
Grau d'enginyeria Informàtica

Bartomeu Antich Luque

12 de gener de 2016

FITXA DEL TREBALL FINAL

Títol del treball:	*Anàlisi de mobilitat d'estudiants Erasmus (Construcció i explotació d'un Data Warehouse)*
Nom de l'autor:	*Jordi Feliu Sobré*
Nom del consultor:	*Bartomeu Antich Luque*
Data de lliurament (mm/aaaa):	*01/2016*
Àrea del Treball Final:	*Magatzems de dades*
Titulació:	*Pla d'Estudis de l'Estudiant*

Resum del Treball (màxim 250 paraules):

L'objectiu del treball és el de construir un magatzem de dades per poder explotar-lo mitjançant eines d'anàlisis multidimensional. Per tal de fer això s'haurà d'aprendre la matèria relacionada amb el magatzems de dades, posar en practica els coneixements de gestió de projectes, aprendre a utilitzar el programari disponible i fer recerca sobre tot allò que no coneguem.

Les dades que disposem per construir el magatzem de dades són de mobilitats d'estudiants d'Erasmus proporcionades per la Unió Europea. Els informes que es volen extreure contindran dades com el nombre de mobilitats per any, edats, universitats amb més mobilitats, etc...

El treball consta de 4 entregues, essent aquest document part de l'entrega final. La primera entrega consta del pla de treball i una anàlisi preliminar de les dades. La segona entrega es centra en l'anàlisi en profunditat de les dades, el disseny del magatzem de dades i el disseny del procés ETL. En la tercera entrega s'implementarà el disseny de l'etapa anterior a la màquina virtual i es podrà veure la importància de la part ETL per la correcta construcció de tot el procés.

Finalment s'obtindrà el producte amb els informes demanats i l'eina ens permetrà fàcilment poder ampliar l'anàlisi des de més punts de vista diferents.

Abstract (in English, 250 words or less):

The aim of the work is to build a data warehouse to exploit it using multidimensional analysis tools. In order to do this, you must learn the material related to data warehouse, use the project management knowledge, learn to use the software available and do research.

The data we have to build the data warehouse is about Erasmus student mobility and is provided by the European Union. The reports we want contain data about the number of mobilities per year, ages, universities with more mobilities, etc ...

The work consists of 4 deliveries, being this document part of final delivery. The first part consists of the work plan and preliminary analysis of the data. The second part focuses on in-depth analysis of the data, the design of the data warehouse and ETL design process. In the third installment is implemented the design stage before the virtual machine and you can see the importance of the ETL for the proper construction of the whole process.

Finally we get the product with the requested reports and the tool easily allows us to expand further analysis from different perspectives.

Paraules clau (entre 4 i 8):

Magatzem de dades, dimensió, fet, ETL, MySQL, Pentaho, MDX, OLAP

Índex

Índex de taules

Índex d'il·lustracions

1. Introducció

1.1 Context i justificació del Treball

Actualment s'usen diferents tecnologies que faciliten la creació, distribució i manipulació de la informació generant volums enormes de dades. Aquesta informació juga un paper molt important en les activitats socials, culturals i econòmiques.

Així doncs, les organitzacions disposen d'una enorme quantitat de dades que provenen de fonts molt diverses, que són emmagatzemades en formats i sistemes d'informació diferents. Aquest fet provoca que l'explotació d'aquestes dades sigui ineficient en l'ajuda a la presa de decisions.

Els magatzems de dades sorgeixen per donar solució al tractament de grans volums de dades. Son una col·lecció de dades que recullen informació de múltiples fonts disperses i que estan orientades al tema, integrades, no volàtils i historiades, organitzades per facilitar l'anàlisi i donar suport a processos d'ajuda a la decisió.

Per tal d'aplicar aquesta solució de magatzem de dades s'aprofita que la Unió Europea disposa d'un portal de dades obertes (https://open-data.europa.eu/es/data) on es pot obtenir una gran varietat de dades de les institucions i organismes de la Unió Europea.

D'aquestes dades s'han considerat d'especial interès les de mobilitat d'estudiants dins el programa Erasmus.

Amb aquestes dades es planteja dissenyar, construir i explotar un magatzem de dades que permeti realitzar una anàlisi en profunditat sobre el moviment d'estudiants en base a diferents eixos d'anàlisi.

1.2 Objectius del Treball

L'objectiu principal és desenvolupar un projecte que permeti ampliar els coneixements adquirits en les assignatures obligatòries de l'àrea de bases de dades. Es tracta d'aprofundir en les bases de dades que donen suport en la presa de decisions a les organitzacions al mateix temps que s'apliquen altres coneixements adquirits al llarg dels Estudis.

S'adquirirà experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible en una base de dades transaccional.

Objectius generals:

- Posar en pràctica els coneixements i competències adquirits durant la titulació.
- Realitzar un projecte passant per totes les seves fases, seleccionant els procediments més adequats per dur-ho a terme.
- Documentar i justificar el desenvolupament i resultat del treball.
- Presentar i defensar el treball realitzat.
- Autoavaluar el treball d'acord amb uns criteris determinats.

Objectius específics:

El TFG ens demana integrar les fonts de dades proporcionades per la Unió Europea amb l'objectiu de realitzar diferents tipus d'anàlisi, per fer-ho necessitarem:

- Comprendre que és un magatzem de dades i quins avantatges ens aporta enfront les bases de dades operacionals.
- Saber fer un disseny multidimensional a partir d'un conjunt de requisits que descriuen una problemàtica donada i aportar noves visions als informes.
- Conèixer els problemes de la integració, transformació i càrrega de dades i saber resoldre-la per a crear un magatzem de dades a partir de múltiples fonts.
- Saber crear un magatzem de dades partint d'un disseny multidimensional i implementar-ho fent servir la tecnologia que es consideri més adequada.
- Conèixer aplicacions i eines per a una òptima explotació del magatzem de dades
- Obtenir els informes requerits.

1.3 Enfocament i mètode seguit

Degut a la naturalesa d'un treball final de grau i el model d'avaluació de la UOC, la metodologia de desenvolupament de software ve imposada des d'un principi a l'enunciat.

Segons les entregues proposades (PAC) el mètode segueix les fases del tipus de desenvolupament en cascada. Aquest desenvolupament és un procés de disseny seqüencial, d'ús freqüent en els processos de desenvolupament de programari, en què el progrés flueix constantment cap avall (com una cascada) a través de les fases de concepció, iniciació, anàlisi, disseny, construcció, proves, producció/realització, i manteniment.

Al punt 1.4 es veurà aquest fet al explicar en que consisteixen les entregues.

1.4 Planificació del Treball

1.4.1 Recursos de maquinari i programari

Els recursos amb els que s'ha comptat per la realització del treball són els següents:

Ordinador personal amb els següents components instal·lats:

- Ofimàtica: Microsoft Office 2013
- Gestor de projectes: GanttProject 2.7.1
- UML: MagicDraw 17.05

Per tal de fer la implementació s'accedeix través d'un enllaç a l'aula a Amazon Elastic Compute Cloud (Amazon EC2) que forma part de la plataforma de còmput en el núvol de l'empresa Amazon.com anomenada Amazon Web Services. EC2 permet llogar computadors virtuals per hores on poder executar aplicacions. El computador virtual disposa de les següents característiques:

- 2 Processadors: Intel Xenó E5-2680 a 2,80GHz v2
- RAM: 3,75GB
- Espai de disc lliure: 38.8GB

L'entorn té instal·lats els següents components:

- Sistema operatiu: Ubuntu 14.04.3 LTS.
- Eines per el disseny de la base de dades: MySQL Workbench 6.3.4.0.
- Suite BI per l'elaboració d'informes: Pentaho Business Analytics 5.0.1 més els visors OLAP Saiku 3.3.2 i CDE 14.12.10.1.
- Eines d'ETL: Component de Pentaho PDI Spoon (Kettle 5.4.0.1).
- Eines Multidimensionals per crear cubs OLAP: Mondrian Schema-Workbench 3.10.0.1.

1.4.2 Tasques a realitzar

Les tasques ha realitzar son:

PAC 1 Pla de Treball i anàlisi preliminar

- Realització del pla de treball: Inclou la justificació del projecte, objectius, requeriments, estructura, fases, riscos i materials.
- Anàlisi preliminar: Inclou l'anàlisi d'informes sol·licitats, fonts de dades, identificar elements DW.

PAC 2 Anàlisi de requeriments i Disseny conceptual i tècnic

- Anàlisi de requeriments final: Requeriments funcionals i no funcionals.
- Disseny del model conceptual.
- Disseny de la BD / Diagrama E-R
- Model multi-dimensional detallat.
- Procés ETL a alt nivell, indicant què fa, com ho fa i perquè ho fa.
- Explicació de com tractar els errors en la càrrega (qualitat de les dades).

PAC 3 Implementació

- Implementació del magatzem de dades
- Implementació del procés ETL.
- Implementació procés OLAP i informes.
- Explicació sobre com s'ha dut a terme el treball demanat.
- Captures de pantalla de tots els informes realitzats i una explicació breu.
- Comentaris rellevants sobre el desenvolupament dut a terme.
- Justificació del compliment dels requisits funcionals.

PAC 4 Memòria i presentació

- Redacció de la memòria segons els resultats obtinguts a les PAC.
- Elaboració de la presentació.
- Informe d'autoavaluació

Al següent quadre es mostren les fites de les 4 entregues:

Lliurables	Contingut	Data d'inici	Data d'entrega
PAC1	Pla de Treball i anàlisi preliminar	19/09/2015	01/10/2015
PAC2	Anàlisi de requeriments i Disseny	02/10/2015	29/10/2015
PAC3	Implementació	30/10/2015	17/12/2015
PAC4	Memòria i Presentació virtual	18/12/2015	12/01/2016
Debat i defensa del projecte	Debat i defensa del projecte	25/01/2016	28/01/2016

Taula 1. Fites

El TFG equival a 12 crèdits ECTS i cada crèdit equival a 25-30 hores. Per tant, s'estima que les hores de dedicació total han de ser entre 300 i 360.

Tenint en compte que des de l'inici de l'assignatura el dia 16/09/05 fins l'entrega final el dia 12/01/16 hi han 119 dies, es calcula una mitjana d'entre 2,5 i 3 hores de treball diari.

No es podrà dedicar les hores calculades tots el dies. S'ajustaran les hores perdudes com s'indica a l'anàlisi de riscos.

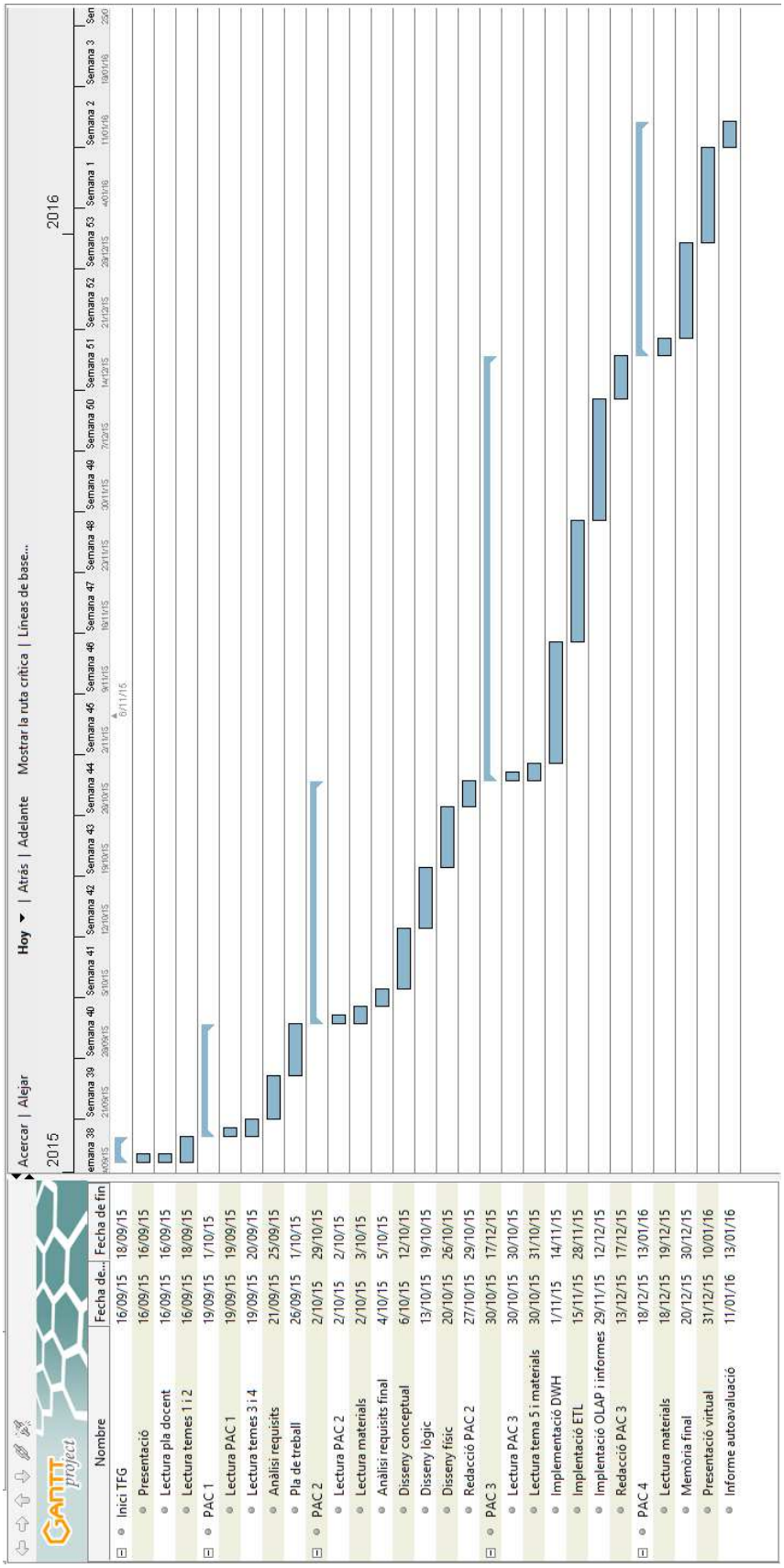
El cronograma proposat pel projecte és el següent:

Tasca	Inici	Fi	Dies
Inici TFG	**16/09/2015**	**18/09/2015**	**3**
Presentació	16/09/2015	16/09/2015	1
Lectura Pla docent	16/09/2015	16/09/2015	1
Lectura tema 1 i 2 materials	17/09/2015	18/09/2015	2
PAC 1	**19/09/2015**	**01/10/2015**	**13**
Lectura PAC 1	19/09/2015	19/09/2015	1
Lectura tema 3 i 4 materials	19/09/2015	20/09/2015	2
Anàlisi requisits	21/09/2015	25/09/2015	5
Pla de treball	26/09/2015	01/10/2015	6
Lliurament PAC 1	**01/10/2015**		**FITA**
PAC 2	**02/10/2015**	**29/10/2015**	**28**
Lectura PAC 2	02/10/2015	02/10/2015	1
Lectura materials necessaris	02/10/2015	03/10/2015	2
Anàlisi requisits final	04/10/2015	05/10/2015	2
Disseny conceptual	06/10/2015	12/10/2015	7
Disseny lògic	13/10/2015	19/10/2015	7
Disseny físic	20/10/2015	26/10/2015	7
Redacció PAC 2	27/10/2015	29/10/2015	3
Lliurament PAC 2	**29/10/2015**		**FITA**
PAC 3	**30/10/2015**	**17/12/2015**	**49**
Lectura PAC 3	30/10/2015	30/10/2015	1
Lectura tema 5 i materials	30/10/2015	31/10/2015	2
Implementació DWH	01/11/2015	14/11/2015	14
Implementació ETL	15/11/2015	28/11/2015	14
Implementació OLAP i informes	29/11/2015	12/12/2015	14
Redacció PAC 3	13/12/2015	17/12/2015	5
Lliurament PAC 3	**17/12/2015**		**FITA**
PAC 4	**18/12/2015**	**12/01/2016**	**26**
Lectura PAC 4	18/12/2015	18/12/2015	1
Lectura materials	18/12/2015	19/12/2015	2
Memòria final	20/12/2015	30/12/2015	11
Presentació virtual	31/12/2015	10/01/2016	11
Informe autoavaluació	11/01/2016	12/01/2016	2
Lliurament PAC 4	**12/01/2015**		**FITA**

Taula 2. Cronograma

S'ha de tenir en compte a partir de la PAC 2 també es dedica un temps a corregir els errors de les PAC anteriors.

Il·lustració 1 amb el diagrama de Gantt corresponent al cronograma anterior:



Il·lustració 1. Diagrama de Gantt

1.4.3 Anàlisi de riscos

L'anàlisi determinarà quins són els factors de risc que potencialment tindran major efecte sobre el projecte i com s'han de gestionar. Els següents quadres mostren els riscos que s'han tingut en compte:

RISC 1	Personal
Event	Augment de la carrega de treball laboral
Probabilitat	Mitja
Impacte	Lliuraments
Mitigació	Dedicar tot el temps possible a la resolució quan la carrega de treball és menor, aprofitar bé els caps de setmana.

RISC 2	Personal
Event	Malaltia personal/familiar
Probabilitat	Baixa
Impacte	Lliuraments
Mitigació	Si la malaltia només afecta uns dies s'aplicarà el pla de mitigació anterior. En cas de que sigui més greu parlar-ho amb el consultor.

RISC 3	Ordinador personal
Event	Fallada Hardware
Probabilitat	Baixa
Impacte	Lliuraments
Mitigació	Reparar l'ordinador i usar el portàtil. Es generen copies de seguretat diàries externes a Dropbox.

RISC 4	Ordinador personal
Event	Fallada Software
Probabilitat	Baixa
Impacte	Lliuraments
Mitigació	Reinstal·lació del software i usar el portàtil si el problema persisteix. Es generen copies de seguretat diàries externes a Dropbox

RISC 5	Telecomunicacions
Event	Fallada fibra òptica / router
Probabilitat	Baixa
Impacte	Lliuraments
Mitigació	Reclamació a la companyia proveïdora. Usar la xarxa mòbil.

RISC 6	Escriptori virtual
Event	Fallada AmazonWS
Probabilitat	Baixa
Impacte	Lliuraments
Mitigació	Instal·lació del software en local.

RISC 7	PAC
Event	Errors en l'execució de les PAC
Probabilitat	Mitja
Impacte	Puntuació treball, pla de treball
Mitigació	Preguntes al consultor. Correcció en la següent PAC. Consultar continguts mínims.

RISC 8	Adaptació al programari
Event	Major dificultat de l'esperada en aprendre a utilitzar programari nou.
Probabilitat	Mitja
Impacte	Lliuraments
Mitigació	Preguntes al consultor. Consulta tutorials/manuals a Internet.

1.5 Breu sumari de productes obtinguts

Els productes obtinguts seran els documents que s'entreguin a cada fita i tot el que es generi a la màquina virtual.

Com s'ha vist al punt anterior la PAC1 contindrà els documents del pla de treball i l'anàlisi preliminar.

La PAC2 consta d'un sol document amb l'anàlisi de requeriments i el disseny conceptual i tècnic.

La PAC3 és la d'implementació, on s'haurà generat el magatzem de dades, el procés ETL de carrega i d'actualització, el cub OLAP i els informes a la maquina virtual. Addicionalment també s'elabora el document on s'explica tot el procés.

Pel que fa a la PAC4 aquest document de memòria en forma part. També contindrà una presentació virtual i l'informe d'autoavaluació.

1.6 Breu descripció dels altres capítols de la memòria

Els següents capítols de la memòria contindran la feina desenvolupada a les entregues PAC2 i PAC3:

- Anàlisi de requeriments i Disseny conceptual i tècnic. Aquest capítol defineix l'arquitectura del procés amb les eines de les que disposem. Es veuen els casos d'us i l'anàlisi de requeriments funcionals i no funcionals. Posteriorment es fa l'anàlisi exhaustiva del fitxers font, es fa el disseny conceptual de l'estrella. Finalment s'acabarà amb el disseny lògic i físic de la base de dades amb els seus corresponents esquemes.

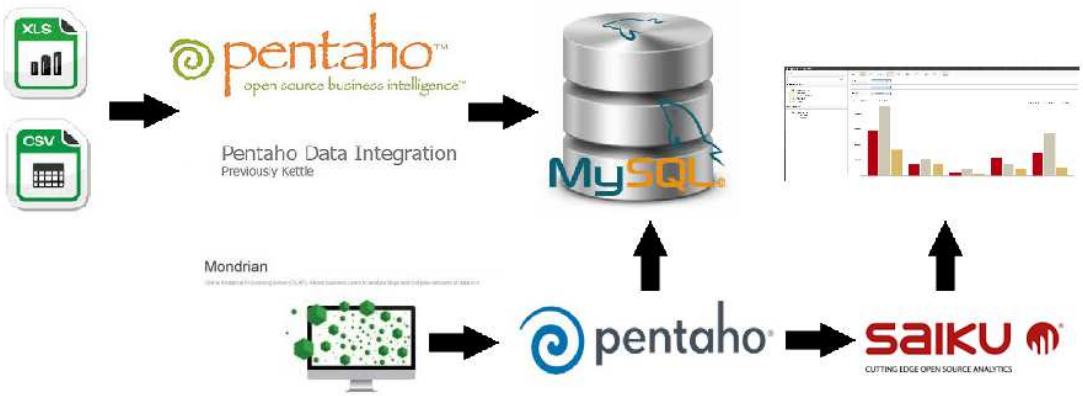
- Implementació. En aquest capítol es mostra la creació de la base de dades. Posteriorment es veu el procés ETL amb les fases de carrega i actualització. Seguidament es genera el cub OLAP per tal de fer l'anàlisi multidimensional i finalment es veuen els informes predefinits i els addicionals.

- Conclusions. A l' últim capítol de la memòria es valoren les lliçons apreses, l'assoliment d'objectius, el seguiment del pla de treball i les línies futures de treball relacionades amb el TFG.

2. Anàlisi de requeriments i Disseny conceptual i tècnic

2.1 Arquitectura del procés

Al punt 1.4 s'han detallat els components instal·lats a la maquina virtual, la il·lustració 2 mostra com s'usen aquests components per tal d'arribar a la solució desitjada:



Il·lustració 2. Arquitectura del procés

Els passos a seguir són els següents:

1. Es parteix de les dades de la Unió Europea, les quals s'han d'analitzar detalladament per poder fer correctament el següent pas.

2. Procés ETL: Un cop fet l'anàlisi detallat i tota la part de disseny es comença amb el procés ETL.

El procés ETL parteix de les dades del pas 1 i finalitza amb el magatzem de dades, que és una base de dades relacional MySQL. El procés ETL s'executa amb el software Spoon-PDI (Kettle).

3. Esquema Mondrian: Esquema XML que permet a l'eina de BI Pentaho accedir a les dades allotjades al magatzem de dades MySQL per a construir els cubs OLAP.

4. Un cop Pentaho tingui carregat l'esquema Mondrian es podrà connectar a MySQL i el motor Mondrian de Pentaho podrà accedir al magatzem de dades mitjançant consultes en llenguatge SQL (prèvia consulta MDX).

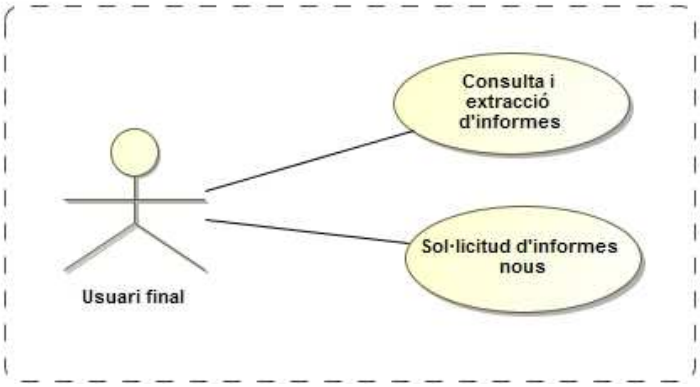
5. L'últim pas és l'objectiu final del treball, l'usuari accedirà als cubs OLAP a través del plugin de consulta MDX anomenat Saiku per veure els informes.

6. Un dels requeriments funcionals és la possibilitat de realitzar actualitzacions dinàmiques anuals del magatzem de dades. Les actualitzacions hauran de passar per un procés ETL específic i afegir-se a les dades ja existents.

2.2 Casos d'ús

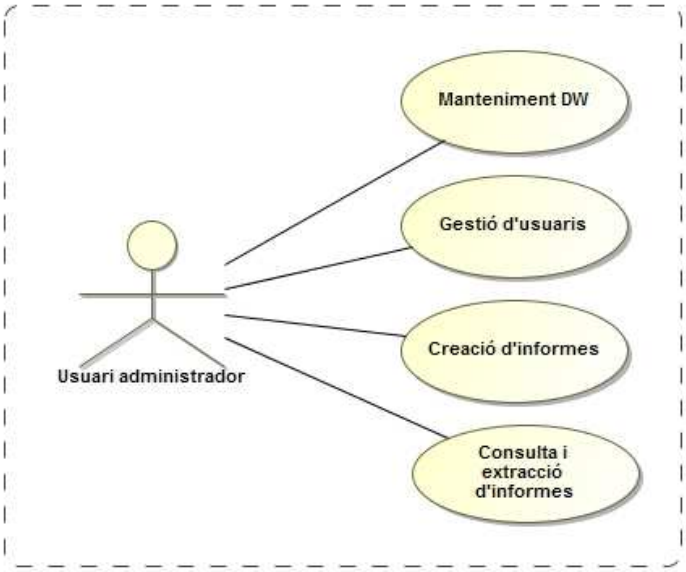
El diagrama de casos d'us ens permetrà veure la relació entre els usuaris i el sistema. L'enunciat del treball no especifica quins rols ha de tenir cada usuari, es defineixen els següents:

- Usuari final: Usuari que consultarà els informes i sol·licitarà de nous a mesura que ho necessiti. La il·lustració 3 mostra el cas descrit:



Il·lustració 3.Cas d'ús usuari final

- Usuari Administrador: És el responsable del manteniment del magatzem de dades incloent els processos d'extracció i càrrega de dades. Te accés a totes les eines relacionades amb els d'informes des de la creació fins l'extracció. Per últim controla la gestió dels usuaris que interactuen amb els sistema. La il·lustració 4 mostra el cas descrit:



Il·lustració 4. Cas d'us usuari administrador

2.3 Requeriments

2.3.1 Requeriments previs

El TFG és una assignatura que està pensada per a ser la darrera que cursi un estudiant en la seva carrera. El TFG és un treball eminentment pràctic i vinculat a l'exercici professional de la informàtica encara que en alguns casos pot ser, o incloure, un treball de recerca.

A més dels coneixements previs generals dels TFG del Grau de Informàtica, és necessari haver cursat les assignatures: Ús de bases de dades i Disseny de bases de dades. També haver treballat amb algun sistema gestor de bases de dades com: PostgreSQL, Oracle, MySQL, MSSQL Server...

2.3.2 Requeriments funcionals

Definiran la funció del sistema. Partint de la informació proporcionada com a resultat s'ha de poder obtenir com a mínim els següents informes:

- 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 aquests informes han de poder ser analitzats comparant els diferents cursos (anys). De partida es disposa de les dades dels cursos 2011-12 i 2012-13.

En l'anàlisi preliminar es va detectar que en el fitxer del curs 2011-12 no hi ha manera d'identificar estudiants i en el fitxer del curs 2012-13 si que n'hi ha però el camp conté suficients errors com per no donar una informació fiable. Per tant hi ha un conflicte en els informes on es demana comptar estudiants.

Realment però el recompte que té un interès pel treball és el recompte de mobilitats més que el d'estudiants. Per tant es proposa la següent modificació dels informes inicials:

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

Com a informes addicionals es proposa obtenir informació de la duració de les mobilitats i dels crèdits estudiats. També que la informació es pugui veure per gènere i tipus de mobilitat.

L' implementació ha de permetre l'actualització dinàmica de totes les dimensions i els fets. Cal poder actualitzar el magatzem de dades amb les dades que es vagin generant en anys consecutius, així com també noves institucions, nacionalitats, etc.

2.3.3 Requeriments no funcionals

Són atributs de qualitat que poden usar-se per jutjar el comportament del sistema. Alguns requeriments desitjables són els següents:

- Rendiment: El temps de resposta en les consultes ha de ser òptim respecte al volum de dades amb el que treballem.
- Disponibilitat, Operativitat: Mesura el percentatge de temps que el sistema esta actiu. Ha de ser la major possible.
- Accessibilitat: Característica d'un sistema que permet que hi interactuïn usuaris amb discapacitats.
- Usabilitat: Mesura la facilitat amb la que els usuaris utilitzen una eina concreta. El sistema ha de ser intuïtiu i fàcil d'usar.
- Estabilitat: És la propietat dels sistemes que tenen un nivell d'errades reduït. S'han de minimitzar les errades de sistema.
- Cost: Ha d'adaptar-se a les característiques del projecte.
- Escalabilitat: El disseny ha de permetre adaptar-se a necessitats futures. L' implementació ha de permetre l'actualització dinàmica de totes les dimensions i els fets. Cal poder actualitzar el magatzem de dades amb les dades que es vagin generant en anys consecutius, així com també noves institucions, nacionalitats, etc
- Mantenibilitat: És la característica que representa la capacitat del sistema per ser modificat en cas de necessitat. El sistema ha de ser fàcil de mantenir.

2.4 Fonts de dades

Es disposa de 6 fitxers proporcionats per la Unió Europea amb diferents formats:

2.4.1 Student mobility_datadictionary.pdf

Diccionari de dades per la mobilitat d'estudiants del 2012-2013. En aquest cas el diccionari de dades és un document PDF que descriu els tipus de mobilitat existents (S,C,P) i els camps de les taules indicant els atributs, definicions, dominis i validacions/restriccions de dades.

Existeixen 3 tipus de mobilitat:

1. Mobilitat només d'estudi (MOBILITYTYPE = 'S'): l'estudiant passa d'una institució del seu país a una institució a un altre país. Tots els camps relacionats amb l'empresa d'acollida han de romandre buits.(HOSTINSTITUTION, COUNTRYCODEOFHOSTINSTITUTION, PLACEMENTENTERPRISE, COUNTRYOFPLACEMENT, ENTERPRISESIZE, TYPEPLACEMENTSECTOR)

2. Mobilitat d'estudi combinada amb pràctiques (MOBILITYTYPE = 'C'): en aquest cas, les dades de les institucions d'origen i destí s'han d'incloure, i és opcional també per incloure informació sobre l'empresa d'acollida al camp COMMENT. Tota la informació sobre la donació, l'ECTS i la durada ha de ser reportades.

3. Mobilitat de pràctiques (MOBILITYTYPE = 'P'): en aquest cas hi ha una institució origen, però el camp HOSTINSTITUTION s'ha de deixar buit.
Si un estudiant va a una institució a fer pràctiques (no per estudiar) la institució ha de ser reportada com a empresa d'acollida (i no s'ha d'informar cap codi EUC per la institució d'acollida).

Hi ha dos modalitats de pràctiques:

Pràctiques organitzades per una institució (codi de mobilitat ERA02 en el camp LLPLINKPROJECT). El camp CONSORTIUMAGREEMENTNUMBER s'ha de deixar buit.

Pràctiques organitzades per un consorci (codi mobilitat ERA04 en el camp LLPLINKPROJECT). El camp CONSORTIUMAGREEMENTNUMBER no pot deixar-se buit.

NOTA: En cas que un estudiant tingui dos períodes de mobilitat en el mateix any, dels tipus 1 i 3 o dels tipus 2 i 3, s'haurà de reportar en diferents registres (línies).

2.4.2 SM_2011_12.csv (24.2 MB)

Fitxer CSV que conté la Mobilitat estudiants Erasmus curs 2011-2012.

Relació de camps (32):

HOMEINSTITUTION: Codi identificador de l'institució origen de l'estudiant. Relacionat amb la taula EUC_Consolidated_Table_2007_2013.xls.

COUNTRYCODEOFHOMEINSTITUTION: País de la institució d'origen. Relacionat amb la taula ISOCountryCodes081507.xls.

AGE: Edat de l'estudiant.

GENDER: Gènere de l'estudiant (M/F).

NATIONALITY: Nacionalitat de l'estudiant. Relacionat amb la taula ISOCountryCodes081507.xls.

SUBJECTAREA: Codi ISCED97 de l'àrea d'estudi. Relacionat amb la taula ISCED97_Erasmus_subject_codes.xls.

LEVELSTUDY: Nivell d'estudi a la institució d'origen (1,2,3,S).

YEARSPRIOR: Número d'anys d'educació superior anteriors a la mobilitat (0-20).

MOBILITYTYPE: Tipus de mobilitat (S,P,C).Valors explicats en Student mobility_datadictionary.pdf

HOSTINSTITUTION: Institució on l'estudiant efectua el període d'Erasmus. Relacionat amb la taula EUC_Consolidated_Table_2007_2013.xls. Nul en cas de MOBILITYTYPE='P'

COUNTRYCODEOFHOSTINSTITUTION: País de la institució de destí. Relacionat amb la taula ISOCountryCodes081507.xls. Nul en cas de MOBILITYTYPE='P'

PLACEMENTENTERPRISE: Nom de la companyia on l'estudiant fa les pràctiques. Camp informat en el cas de MOBILITYTYPE='P'

COUNTRYOFPLACEMENT: País on es desenvolupen les pràctiques. Relacionat amb la taula ISOCountryCodes081507.xls. Camp informat en el cas de MOBILITYTYPE='P'

ENTERPRISESIZE: Mida de l'empresa (L,M,S) . Camp informat en el cas de MOBILITYTYPE='P'

TYPEPLACEMENTSECTOR: Tipus de sector. Camp informat en el cas de MOBILITYTYPE='P'

LENGTHSTUDYPERIOD: Mesos que l'estudiant ha estat fora. Entre 0.00 i 13.5 a increments de 0.25. Per MOBILITYTYPE=P el camp valdrà 0. Vigilar valors <3.

LENGTHPLACEMENT: Mesos que l'estudiant ha estat fora en pràctiques. Entre 0.00 i 13.5 a increments de 0.25. Per MOBILITYTYPE<>P el camp valdrà 0. Vigilar valors <3. Si LEVELSTUDY = 'S' llavors el camp ha de ser >=2.00.

SHORTDURATION: Raó per la qual l'estudiant ha participat menys de 3 mesos en el programa.('',T,X). Si MOBILITYTYPE<>P i LENGTHSTUDYPERIOD<3 el camp no pot estar buit.

STUDYSTARTDATE: Data en que el període d'estudi comença. Ha de ser major que juny 2011. A vegades comenta amb majúscula i a vegades no. Camp buit si MOBILITYTYPE=P.

PLACEMENTSTARTDATE: Data en que el període de pràctiques comença. Ha de ser major que juny 2011. A vegades comenta amb majúscula i a vegades no. Camp buit si MOBILITYTYPE<>P.

CONSORTIUMAGREEMENTNUMBER: Número d'identificació si les pràctiques són administrades per consorci.

ECTSCREDITSSTUDY: Número de crèdits ECTS. Entre 0 i 90. Vigilar majors de 60. Si MOBILITYTYPE=P llavors valdrà 0.

ECTSCREDITSPLACEMENT: Número de crèdits ECTS de pràctiques . Entre 0 i 90. Vigilar majors de 60. Si MOBILITYTYPE<>P llavors valdrà 0.

TOTALECTSCREDITS: Crèdits totals. Suma dels dos camps anteriors.

SNSUPPLEMENT: Beca rebuda per necessitats especials. >=0. Vigilar <=10000.

TAUGHTHOSTLANG: Aprenentatge realitzat amb el llenguatge de destinació (Y/N).

LANGUAGETAUGHT: Llenguatge amb el que l'estudiant ha realitzat l'aprenentatge. Relacionat amb la taula ISOCountryCodes081507.xls. 'XX' si no apareix.

LINGPREPARATION: Preparació lingüística de l'estudiant. (EC,HS,HM,NN).

STUDYGRANT: Total de beca excloent el SNSUPPLEMENT. 0 si MOBILITYTYPE=P.

PLACEMENTGRANT : Total de beca de pràctiques excloent el SNSUPPLEMENT. 0 si MOBILITYTYPE<>P.

PREVIOUSPARTICIPATION: Indica si l'estudiant ja ha realitzat un Erasmus anterior i amb quin tipus de mobilitat (N/S/P/M).

QUALIFICATIONATHOST: Tipus de qualificació rebuda (D,J,O,N).

Observacions:

- La taula conté 252.828 files incloent la capçalera.
- La taula inclou 32 camps de capçalera dels 45 definits en el diccionari de dades.

- Camps rellevants per generar els informes:
    - HOMEINSTITUTION: 3.189 registres diferents del quals 31 no apareixen al fitxer EUC_Consolidated_Table_2007_2013.xls:

      A LINZ19, B ARLON08, B MONS02, B MONS05, B MONS11, B NAMUR10, CH BRUGG02, D KOLN03, D ZITTAU02, DK ARHUS08, DK RANDERS04, DK XX, E BENLAMA01, E SAB-SEB13, E TENERIF05, EE TALLINN19, F NANCY01, F PARIS098, F TOULOUS123, F XX, HU BUDAPES04, HU BUDAPES30,HU XX, N BEKKEST01, N OSLO23, NL LEEUWAR02, NL VELP03, PL WARSZAW43,PL WARSZAW50, RO BAIA-MA01,UK LONDON049

    - HOSTINSTUTION: 2370 registres diferents contant un nul del quals 45 no apareixen al fitxer EUC_Consolidated_Table_2007_2013.xls:

      A WIENER04, A XX, B ARLON08, B MONS02, B MONS05, B NAMUR10, BEDE XX, BEFR XX, CH BRUGG02, CH XX, CZ XX, D KOLN03, D XX, DK ARHUS08, DK HADERSL02, DK RANDERS04, DK XX, E TENERIF05, E XX, F CERGY05, F PARIS098, F PARIS236, F TOULOUS113, F TOULOUS35, F XX, HU BUDAPES04, HU BUDAPES30, HU BUDAPES52, HU XX, I XX, IS XX, LT XX, LUX XX, N BEKKEST01, N OSLO23, N XX, NL LEEUWAR02, NL VELP03, NL XX, P XX, PL WARSZAW43, PL XX, RO BAIA-MA01, S XX, SF XX, SK XX, TR XX, UK COLWYN01, UK EDINBUR04, UK LONDON049, UK XX

    - PLACEMENTENTERPRISE: 33829 registres incloent un nul. Majoritàriament en majúscules però la mateixa empresa pot estar descrita de maneres diferents. No es disposa d'una taula per verificar-les.

    - NATIONALITY: 35 registres. Tots es troben a la taula ISOCountryCodes081507 excepte 'XX'.

    - GENDER: F=153468 registres, M=99359 registres.

    - MOBILITYTYPE: C=438 registres, P=48083 registres i S=204306 registres.

    - COUNTRYCODEOFHOMEINSTITUTION: 35 registres, 3 dels quals no apareixen a la taula ISOCountryCodes081507:

      BEDE, BEFR, BENL

      S'han de corregir per 'BE'.

    - COUNTRYCODEOFHOSTINSTITUTION: 36 registres incloent un en blanc, 3 dels quals no apareixen a la taula ISOCountryCodes081507:

      '',BEDE, BEFR, BENL

      S'han de corregir per 'BE'.

    - COUNTRYOFPLACEMENT: 35 registres incloent un en blanc, 1 dels quals no apareixen a la taula ISOCountryCodes081507:

      BEFR

      S'ha de corregir per 'BE'.

- SUBJECTAREA: 143 registres numèrics, tots apareixen a la taula ISCED97_Erasmus_subject_codes, però sen dupliquen 2 ja que a la taula hi ha els codis caràcter '1','01','8','08' que al transformar-los a numèric resulten 1 i 8 amb dos descripcions diferents.

- Parella HOMEINSTITUTION – COUNTRYCODEOFHOMEINSTITUTION: A banda dels casos anteriorment detectats de les 31 institucions i dels 3 països que s'han de corregir per 'BE' és possible que hi hagin parelles institució-país que no siguin correctes:

  F LYON104, F MARSEIL84, F MARSEIL94, F NANCY43, F PARIS037, F RENNES49 han de tenir informat el codi 'FR'

  G KALAMAT01 ha de tenir informat el codi 'GR'

- Parella HOSTINSTITUTION – COUNTRYCODEOFHOSTINSTITUTION: A banda dels casos anteriorment detectats de les 45 institucions i dels 3 països que s'han de corregir per 'BE' és possible que hi hagin parelles institució-país que no siguin correctes:

  G KALAMAT01 ha de tenir informat el codi 'GR'

2.4.3 SM_2012_13.csv (43.6 MB)

Fitxer CSV que conté la Mobilitat estudiants Erasmus curs 2012-2013

Relació de camps (34):

STUDENT_ID: Identificador únic per cada estudiant (DNI).

ID_MOBILITY_CDE: Clau primària generada automàticament.

HOME_INSTITUTION_CDE: Codi identificador de la institució origen de l'estudiant. Relacionat amb la taula EUC_Consolidated_Table_2007_2013.xls.

HOME_INSTITUTION_CTRY_CDE: País de la institució d'origen. Relacionat amb la taula ISOCountryCodes081507.xls.

STUDENT_AGE_VALUE: Edat de l'estudiant

STUDENT_GENDER_CDE: Gènere de l'estudiant (M/F).

STUDENT_NATIONALITY_CDE: Nacionalitat de l'estudiant. Relacionat amb la taula ISOCountryCodes081507.xls.

STUDENT_SUBJECT_AREA_VALUE: Codi ISCED97 de l'àrea d'estudi. Relacionat amb la taula ISCED97_Erasmus_subject_codes.xls

STUDENT_STUDY_LEVEL_CDE: Nivell d'estudi a la institució d'origen (1,2,3,S).

NUMB_YRS_HIGHER_EDUCAT_VALUE: Número d'anys d'educació superior anteriors a la mobilitat (0-20).

MOBILITY_TYPE_CDE: Tipus de mobilitat (S,P,C).Valors explicats en Student mobility_datadictionary.pdf

HOST_INSTITUTION_CDE: Institució on l'estudiant efectua el període d'Erasmus. Relacionat amb la taula EUC_Consolidated_Table_2007_2013.xls. Nul en cas de MOBILITYTYPE='P'

HOST_INSTITUTION_COUNTRY_CDE: País de la institució de destí. Relacionat amb la taula ISOCountryCodes081507.xls. Nul en cas de MOBILITYTYPE='P'

PLACEMENT_ENTERPRISE_VALUE: Nom de la companyia on l'estudiant fa les pràctiques. Camp informat en el cas de MOBILITYTYPE='P'.

PLACEMENT_ENTERPRISE_CTRY_CDE: País on es desenvolupen les pràctiques. Relacionat amb la taula ISOCountryCodes081507.xls. Camp informat en el cas de MOBILITYTYPE='P'.

PLACEMENT_ENTERPRISE_SIZE_CDE: Mida de l'empresa (L,M,S) . Camp informat en el cas de MOBILITYTYPE='P'.

TYPE_PLACEMENT_SECTOR_VALUE: Tipus de sector . Camp informat en el cas de MOBILITYTYPE='P'.

LENGTH_STUDY_PERIOD_VALUE: Mesos que l'estudiant ha estat fora. Entre 0.00 i 13.5 a increments de 0.25. Per MOBILITYTYPE=P el camp valdrà 0. Vigilar valors <3.

LENGTH_PLACEMENT_VALUE: Mesos que l'estudiant ha estat fora en pràctiques. Entre 0.00 i 13.5 a increments de 0.25. Per MOBILITYTYPE<>P el camp valdrà 0. Vigilar valors <3. Si LEVELSTUDY = 'S' llavors el camp ha de ser >=2.00.

SHORT_DURATION_CDE: Raó per la qual l'estudiant ha participat menys de 3 mesos en el programa.('',T,X). Si MOBILITYTYPE<>P i LENGTHSTUDYPERIOD<3 el camp no pot estar buit.

STUDY_START_DATE: Data en que el període d'estudi comença. Ha de ser major que juny 2012. A vegades comenta amb majúscula i a vegades no. Camp buit si MOBILITYTYPE=P.

PLACEMENT_START_DATE: Data en que el període de pràctiques comença. Ha de ser major que juny 2012. A vegades comenta amb majúscula i a vegades no. Camp buit si MOBILITYTYPE<>P.

CONSORTIUM_AGREEMENT_NUMBER: Número d'identificació si les pràctiques són administrades per consorci.

ECTS_CREDITS_STUDY_AMT: Número de crèdits ECTS. Entre 0 i 90. Vigilar majors de 60. Si MOBILITYTYPE=P llavors valdrà 0.

ECTS_CREDITS_PLACEMENT_AMT: Número de crèdits ECTS de pràctiques . Entre 0 i 90. Vigilar majors de 60. Si MOBILITYTYPE<>P llavors valdrà 0.

TOTAL_ECTS_CREDITS_AMT: Crèdits totals. Suma dels dos camps anteriors.

SPECIAL_NEEDS_SUPPLEMENT_VALUE: Beca rebuda per necessitats especials. >=0. Vigilar <=10000.

TAUGHT_HOST_LANGUAGE_CDE: Aprenentatge realitzat amb el llenguatge de destinació (Y/N).

LANGUAGE_TAUGHT_CDE: Llenguatge amb el que l'estudiant ha realitzat l'aprenentatge. Relacionat amb la taula ISOCountryCodes081507.xls. 'XX' si no apareix.

LINGUISTIC_PREPARATION_CDE: Preparació lingüística de l'estudiant. (EC,HS,HM,NN).

STUDY_GRANT_AMT: Total de beca excloent el SNSUPPLEMENT. 0 si MOBILITYTYPE=P.

PLACEMENT_GRANT_AMT: Total de beca de pràctiques excloent el SNSUPPLEMENT. 0 si MOBILITYTYPE<>P.

PREVIOUS_PARTICIPATION_CDE: Indica si l'estudiant ja ha realitzat un Erasmus anterior i amb quin tipus de mobilitat (N/S/P/M).

QUALIFICATION_AT_HOST_CDE: Tipus de qualificació rebuda (D,J,O,N).

Observacions:

- La taula conté 267548 files incloent la capçalera.
- La taula inclou 34 camps de capçalera dels 45 definits en el diccionari de dades. Els camps de capçalera també canvien de nom respecte al diccionari de dades.
- Hi han dos camps de diferencia respecte la taula de l'any anterior: STUDENT_ID i ID_MOBILITY_CDE.
- El camp STUDY_START_DATE conté mesos que comencen amb majúscula i altres amb minúscula.
- La taula te un problema amb el tractament de buits i nuls ja que els hi assigna '???' o bé '? Unknown ?'.
- Camps rellevants per generar els informes :
    - HOME_INSTITUTION_CDE: 3281 registres diferents del quals 38 no apareixen al fitxer EUC_Consolidated_Table_2007_2013.xls:

      BEFR XX, CH BRUGG02, CZ Praha32, D Aachen01, D Aachen02, D Berlin02, D Berlin03, D Berlin18, D Berlin21, D Bochum02, D FRANKFu03, D Frankfu10, D Hamburg01, D Hamburg03, D Hamburg05, D Hamburg06, D Hamburg11, D Hamburg14, D Hamburg17, D Hannove03, D Hannove09, D Idstein01, D Isny01, D Mainz05, D Munster05, D Saarbru08, D Speyer02, D Wedel01, D Witten02, D Wurzbur02, DK XX, E SAB-SEB13, F XX, NL Utrecht34, Nl MAASTRI01, P Lisboa08, PL XX, e palma17

    - HOST_INSTITUTION_CDE: 2491 registres diferents contant un nul del quals 74 no apareixen al fitxer EUC_Consolidated_Table_2007_2013.xls:

      ???, A Graz09, A WIENER04, A XX, BEFR XX, BENL XX, BG Plovdiv01, BG Plovdiv02, BG Varna04, CH BRUGG02, CH Bern01, CH ST.Gall08, CH XX, CH Zurich20, CY Nicosia14, CY XX, CZ Praha07, CZ XX, D Potsdam01, D XX, DK Kobenha58, DK Kopenha10, DK XX, DK xx, E BENLAMA01, E Madrid03, E XX, F Bordeau03, F CERGY05, F Lyon25, F Lyon58, F Marseil51, F Nantes01, F Paris011, F Paris104, F Paris117, F TOULOUs03, F Toulous48, F XX, G Thessal01, G XX, HU XX, I Torino01, I XX, IRLDublin01, IS Reykjav06, LT Kaunas02, LT XX, N XX, NL Amsterd57, NL Groning01, NL Groning03, NL XX, P Lisboa02, P Lisboa46, P XX, PL Opole02, PL Poznan08, PL XX, Pl WARSZAW14, RO Clujnap06, S Stockho10, S XX, SF Vantaa06, SF XX, SI Ljublja01, TR Ankara03, TR Istanbu25, TR XX, UK Edinbur01, UK XX, UK sheffie02, Uk LONDON062, e alicant01

    - PLACEMENT_ENTERPRISE_VALUE: 39770 registres incloent un nul. Registres en majúscules, minúscules i la mateixa empresa pot estar descrita de maneres diferents. No es disposa d'una taula per verificar-les.

    - STUDENT_GENDER_CDE: F=162962 registres, M=104585 registres.

- MOBILITY_TYPE_CDE: C=476 registres, P=55552 registres, S=211519 registres.

- STUDENT_NATIONALITY_CDE: 46 registres. Tots es troben a la taula ISOCountryCodes081507 menys:

  De, Dk, Es, Hu, Nl, No, Pl, XX, de, es, pt, xx

  S'han de transformar a majúscules.

- HOME_INSTITUTION_CTRY_CDE: 35 registres, 3 dels quals no apareixen a la taula ISOCountryCodes081507:

  BEDE, BEFR, BENL

  S'han de corregir per 'BE'.

- HOST_INSTITUTION_COUNTRY_CDE: 36 registres incloent un en blanc, 3 dels quals no apareixen a la taula ISOCountryCodes081507:

  ???, BEDE, BEFR, BENL

  S'han de corregir per 'BE'.

- PLACEMENT_ENTERPRISE_CTRY_CDE: 37 registres incloent un en blanc, 1 dels quals no apareixen a la taula ISOCountryCodes081507:

  ???, BENL

  S'ha de corregir per 'BE'.

- STUDENT_SUBJECT_AREA_VALUE: 141 registres numèrics, tots apareixen a la taula ISCED97_Erasmus_subject_codes, però sen dupliquen 2 ja que a la taula hi ha els codis caràcter '1','01','8','08' que al transformar-los a numèric resulten 1 i 8 amb dos descripcions diferents.

- Parella HOME_INSTITUTION_CDE – HOME_INSTITUTION_CTRY_CDE: A banda dels casos anteriorment detectats de les 38 institucions i dels 3 països que s'han de corregir per 'BE' és possible que hi hagin parelles institució-pais que no siguin correctes:

  G KALAMAT01 ha de tenir informat el codi 'GR'

- Parella HOST_INSTITUTION_CDE – HOST_INSTITUTION_CTRY_CDE: A banda dels casos anteriorment detectats de les 74 institucions i dels 3 països que s'han de corregir per 'BE' és possible que hi hagin parelles institució-pais que no siguin correctes:

  G KALAMAT01 ha de tenir informat el codi 'GR'

2.4.4 ISOCountryCodes081507.xls (50 KB)

Fitxer .xls que conté la relació entre codi i el nom de la nacionalitat.

Camps (2):

CODE: Codi del país. Exemple: 'ad'.

COUNTRY: Nom del país Exemple: 'Andorra'.

Observacions:

- 253 files incloent la capçalera.
- Cal fer notar que el codi de país en aquesta taula esta codificat en lletres minúscules a diferencia de les altres taules. S'ha de transformar el codi de país a majúscules per adequar-lo a la resta de taules.
- De la revisió de dades s'observa que el camp CODE són dominis d'Internet i per tant alguns no correspondran a cap país. A part d'aquesta observació no hi han camps nuls.
- Només es necessiten els països de la unió europea que participen al programa Erasmus.

2.4.5 EUC_Consolidated_Table_2007_2013.xls (1069 KB)

Llistat amb informació de les institucions.

Camps (7):

COUNTRY: Codi del país. Exemple: 'FR'.

CHARTER TYPE CODE: Codi tipus de carta (EUC,EUCP,EUCX).

ORGANISATION NAME: Nom de l'organització.

ERASMUS CODE: Codi identificador de la institució.

STREET: Carrer.

POSTCODE: Codi postal.

CITY: Ciutat.

Observacions:

- 4919 files incloent la capçalera.
- Hi han 5 files amb el carrer buit i 1 sense codi postal.
- El codi Erasmus 'A WIENER01' apareix dues vegades. Buscant per internet el registre correcte és el que pertany a al ciutat de Wiener Neustadt.

- Les ciutats estan codificades amb diferents formats, majúscules, minúscules, idiomes i descripcions diferents.

- Dels 35 codis de països tots apareixen a la taula de països menys 'EL'. Es tracta del cas detectat anteriorment que pertany a 'GR'.

2.4.6 ISCED97_Erasmus_subject_codes.xls

Llistat codis i atributs àrees de coneixement / especialitzacions (Equivalent ISCED97 clau SUBJECTAREA). La pestanya necessària és ISCED97 codes

Camps (4):

Code: Codi ISCED97. Exemple: '010'.

Description: Nom de l'àrea. Exemple: 'Basic/broad, general programmes'.

Equivalent ERASMUS Short Code (Only for reference): Exemple: 62.

Observacions:

- 146 files incloent la capçalera.

- Hi ha 4 camps del code que entren en conflicte ja que hauran de ser numèrics per adaptar-se a les altres taules:

01	Basic/broad, general programmes
1	Education
8	Services
08	Literacy and numeracy

Aquests camps al passar-los a numèric faran que apareguin dos descripcions pel codi 1 i 2 descripcions pel codi 8 cosa que provocarà duplicats. A la següent taula es veu que hi ha descriptius repetits:

080	Literacy and numeracy
010	Basic/broad, general programmes

Per tant es poden eliminar el registres següents:

08	Literacy and numeracy
01	Basic/broad, general programmes

I no perdre informació.

Si es mira a les taules de mobilitat s'observen 1058 registres amb els codis anteriors per 2011-12 i 1092 registres per 2012-13. Aquest registres suposen menys d'un 0.5% del total de registres cosa que suposa un baix impacte en el cas de pedre registres.

2.5 Disseny

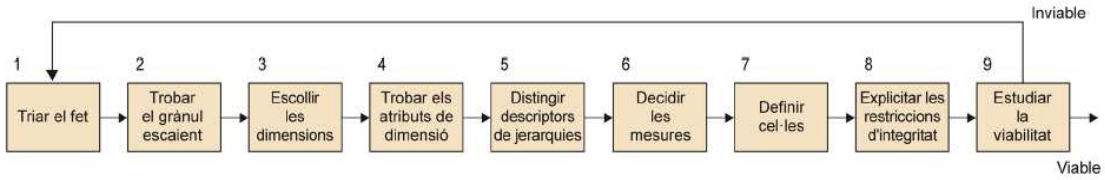
2.5.1 Disseny Conceptual

Un cop s'han analitzat les fonts de dades es proposa un disseny d'alt nivell que serà independent del SGBD que utilitzem.

La característica més important dels magatzems de dades és la multidimensionalitat que consisteix en concebre les dades que es volen analitzar en termes de fets i dimensions d'anàlisi, de manera que es poden situar en un espai n-dimensional.

Els principals elements d'un esquema multidimensional són, d'una banda, les Dimensions, els Nivells i els Descriptors; i de l'altra, simètricament els Fets, les Cel·les i les Mesures.

La il·lustració 5 mostra el cas basic per identificar els elements DW, la metodologia per a fer el disseny conceptual d'una estrella:



Il·lustració 5. Metodologia per dissenyar una estrella
Font: Alberto Abelló Gamazo. Disseny multidimensional. PID_00189746 UOC

2.5.1.1 Triar el fet

El Fet representa el tema objecte d'anàlisi, un conjunt d'esdeveniments amb dades numèriques associades.

En el cas tractat el fet és la mobilitat dels estudiants d'Erasmus. Es designarà la taula com a f_mobilitat.

2.5.1.2 Trobar el grànul escaient

El grànul és l'individu últim que es vol analitzar, la Cel·la més petita que es vol tenir disponible. Triar un grànul massa gran representa perdre informació. Però triar-lo massa petit pot representar malbaratar espai o arribar a fer inviable el projecte per excés de dades.

El grànul pel nostre fet:

Mobilitats per any, nacionalitat, gènere, àrea d'estudi, tipus de mobilitat, universitat emissora i entitat de destí.

2.5.1.3 Escollir les dimensions

Les dimensions representen el punt de vista des del qual es poden analitzar les dades. El grànul anterior determina un primer conjunt de Dimensions. S'han d'afegir a aquest els altres punts de vista que es vulguin utilitzar en l'anàlisi:

- Temps (temps): Cursos dels quals es disposa de les dades de mobilitat.
- Nacionalitat (d_nacionalitat): Nacionalitat de l'alumne.
- Gènere (d_genere): Gènere de l'alumne.
- Mobilitat (d_mobilitat): Tipus de mobilitat.
- Àrea de coneixement (d_coneixement)
- Entitat (d_entitat): Depenent del tipus de mobilitat l'alumne pot tenir com a destí una universitat o a una empresa. Aquesta dimensió tindrà un doble rol per l'entitat emissora i la de destí.

2.5.1.4 Trobar els atributs de dimensió

S'han de seleccionar els atributs que es creu que puguin ser útils per a seleccionar, agrupar o simplement posar com a capçalera dels informes.

Els atributs han d'estar definits sobre un domini discret, ser descriptius, fàcils de recordar i entenedors a primer cop d'ull. Analitzant els informes requerits necessitem:

- d_temps: Any
- d_nacionalitat: codi, nacionalitat
- d_genere: gènere
- d_tipus_mobilitat: tipus
- d_coneixement: codi, àrea
- d_entitat: codi, entitat, país

2.5.1.5 Distingir descriptors de jerarquies

D'entre els atributs que hi ha en una Dimensió, s'en distingeixen de dos tipus: els que s'utilitzaràn per a agrupar i els que serviran simplement per a seleccionar.

De la dimensió d'entitats de destí es necessitarà distingir entre les universitats i les empreses, però les empreses no tenen codi per tant es definirà un codi únic per les empreses.

2.5.1.6 Decidir les mesures

Les Mesures són atributs numèrics normalment additius. Les mesures contindran les sumes dels atributs que es necessiten agrupades per grànul. Els càlculs per obtenir les mitjanes (la mesura de mobilitats serà el divisor) i els percentatges es deixen per la part final de construcció dels informes.

- mobilitats: Numèric, suma de les mobilitats.
- edat: Numèric, suma d'edat d'estudiants.
- beques: Numèric, suma de les beques.
- credits: Numèric, suma dels crèdits.
- durada: Numèric, suma de les durades.

2.5.1.7 Definir les cel·les

Les cel·les definides correspondran amb els fets identificats:

- Mobilitat dels estudiants d'Erasmus
    - mobilitats
    - edat
    - beques
    - credits
    - durada

2.5.1.8 Explicitar les restriccions d'integritat

Un cop es tenen totes les Mesures, Cel·les i Nivells, tan sols queda expressar les restriccions d'integritat corresponents. Dels diferents conjunts de Nivells que defineixin espais en què es puguin col·locar les instàncies d'una Cel·la s'en diuen Bases.

La cel·la mobilitat es definirà a partir de l'any, nacionalitat, gènere, àrea d'estudi, tipus de mobilitat, universitat emissora i entitat de destí.

Pel que fa a les restriccions d'agregació, totes les operacions de l'estrella són compatibles, disjuntes i completes. Per tant no hi ha cap problema per efectuar operacions d'agregació o transitivitat.

2.5.1.9 Estudiar la viabilitat

S'ha d'estimar l'espai que ocuparan les dades. El mètode més realista és mirar les dades que contenen els sistemes operacionals per a saber quantes contindrà en el nostre sistema d'anàlisi.

Per a implementar l'estrella dissenyada es necessita una taula per al Fet (en què cada fila representa una cel·la de l'espai multidimensional) i una taula més per a cadascuna de les Dimensions. Les jerarquies d'agregació queden implícites en els valors dels atributs de les taules de Dimensió. No s'expliciten amb taules diferents.

Es pot considerar només el que ocuparà emmagatzemar les instàncies del Fet. La taula del Fet serà d'ordres de magnitud més gran que qualsevol de les taules de Dimensió. Ocuparà més d'un 95% de l'espai utilitzat per l'estrella.

Les taules de Dimensió es creen amb les dades a dins i molt rarament canvien. Només de tant en tant s'afegeix una nova fila o es canvia el valor d'una que ja hi havia (mai no s'esborren files). En la taula de Fets s'insereixen files massivament de manera regular i només té modificacions si s'ha comès un error durant la inserció (tampoc no s'esborra mai res).

Per tal de reduir la grandària la taula de Fet estarà lligada per claus foranes amb les taules de Dimensió. Cada clau forana apunta de la taula del Fet cap a una de les taules de les Dimensions. Al costat de les claus foranes, la taula del Fet conté les Mesures, mentre que les taules de Dimensió contenen els Descriptors.

Com a clau primària de la taula del Fet es tindran els atributs corresponents a una de les Bases de la Cel·la atòmica. La resta de Bases de la Cel·la donaran lloc a claus alternatives. En qualsevol cas, tant la clau primària com les alternatives seran subconjunts del conjunt de claus foranes que apunten cap a les taules de Dimensió.

Com que els substituts ocuparan menys espai que els atributs identificadors de la mateixa taula (per exemple, un DNI ocupa vuit caràcters, mentre que un RowID només quatre bytes), utilitzant-los reduïm la grandària de les columnes que formen la clau primària de les taules de Dimensió. Consegüentment, també reduïm la grandària de la clau primària de la taula del Fet, perquè se sap que sempre està formada per claus foranes que apunten a les claus primàries de les taules de Dimensió. A més a més, també serveix per a evitar problemes si es modifiquen els identificadors en els sistemes operacionals.

En aquest cas es pot fer una aproximació abans de tractar les dades al procés ETL comptant amb els tipus de dada que ofereix MySQL:

Pel curs 2011-12 s'obtenen 178.108 combinacions possibles amb els atributs que es necessiten i per 2012-13 s'obtenen 190.201.

Sumant aquests últims càlculs s'obtenen 368.309 possibles cel·les.

Un cop se saben quantes cel·les tindrem, s'ha de calcular el nombre de bytes que ocuparà cada cel·la:

Identificadors:

idTemps: tinyint 1 byte
idNacionalitat: tinyint 1 byte
idConeixement: tinyint 1 byte
idGenere: tinyint 1 byte
idTipus: tinyint 1 byte
idUniversitatE: smallint 2 bytes
idEntitatR: smallint 2 bytes

Mesures:

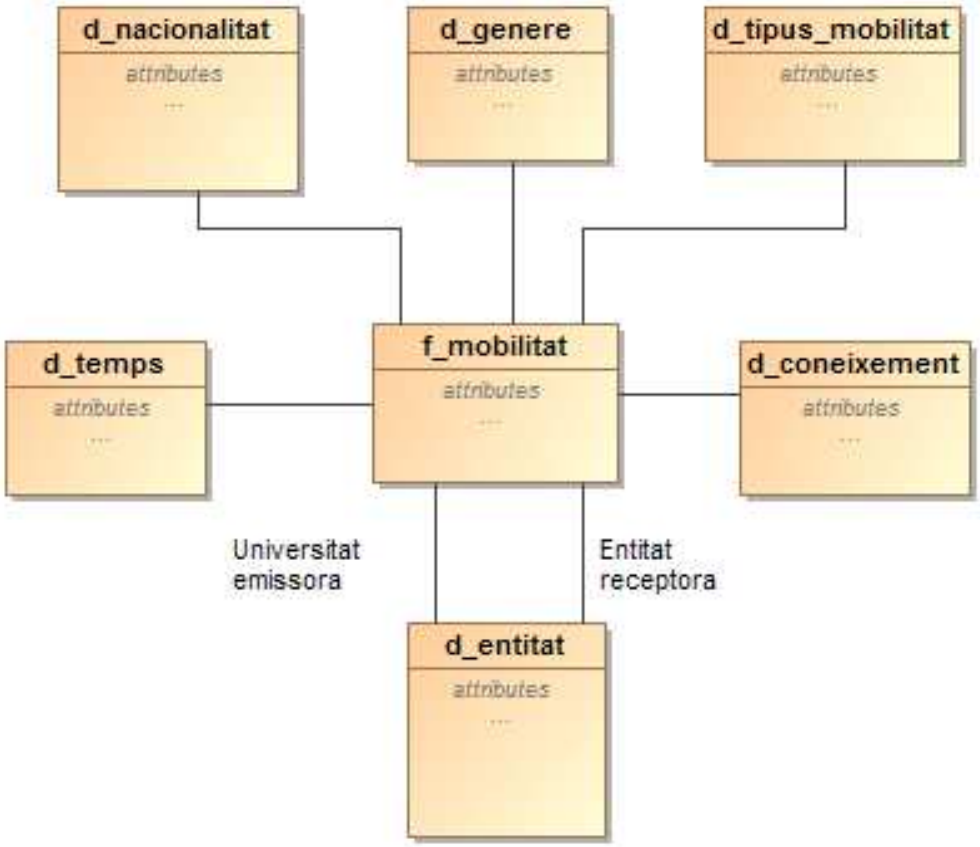
s_mobilitats: smallint 2 bytes
s_edat: smallint 2 bytes
s_beques: float 4 bytes
s_credits: float 4 bytes
s_durada: float 4 bytes

S'obtenen un total de 25 bytes. Multiplicant les cel·les pel nombre de bytes obtenim:

25*368.309 = 9.207.725 bytes = 9.21 Mb

Que és una mida raonable.

La il·lustració 6 mostra el model conceptual de l'estrella dissenyada:



Il·lustració 6. Model conceptual d'estrella

2.5.2 Disseny Tècnic

2.5.2.1 Disseny lògic

El disseny lògic com s'ha vist en l'apartat anterior serà d'estrella, a partir d'aquest es poden definir:

- Fet:
    - f_mobilitat(idTemps, idNacionalitat, idConeixement, idGenere, idTipus, idUniversitatE, idEntitatR, s_mobilitats, s_edat, s_beques, s_credits, s_durada)
- Dimensions:
    - d_temps(rowid, any)
    - d_nacionalitat (rowid, codi, nacionalitat)
    - d_genere(rowid, genere)
    - d_tipus_mobilitat(rowid, tipus)
    - d_coneixement(codi, area)
    - d_entitat(rowid, codi, entitat, país)

2.5.2.2 Disseny físic

El disseny físic de la base de dades conté la descripció de les taules especificant els índex, restriccions i els camps (tipus, mida, claus i restriccions). Seguidament es defineixen les taules pel fet i les dimensions:

f_mobilitat							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
idTemps	tinyint	1	X	X		X	
idNacionalitat	tinyint	1	X	X		X	
idConeixement	tinyint	1	X	X		X	
idGenere	tinyint	1	X	X		X	
idTipus	tinyint	1	X	X		X	
idUniversitatE	smallint	2	X	X		X	
idEntitatR	smallint	2	X	X		X	
s_mobilitats	smallint	2				X	
s_edat	smallint	2				X	
s_beques	float	4				X	
s_credits	float	4				X	
s_durada	float	4				X	

Taula 3. Fet f_mobilitat

d_temps							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
rowId	tinyint	1	X			X	X
any	tinyint	1			X	X	

Taula 4. Dimensió d_temps

d_nacionalitat							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
rowid	tinyint	1	X			X	X
codi	varchar	2			X	X	
nacionalitat	varchar	15			X	X	

Taula 5. Dimensió d_nacionalitat

d_genere							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
rowid	tinyint	1	X			X	X
genere	varchar	1			X	X	
descripcio	varchar	7			X	X	

Taula 6. Dimensió d_genere

d_tipus_mobilitat							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
rowid	tinyint	1	X			X	X
tipus	varchar	1			X	X	
descripcio	varchar	19			X	X	

Taula 7. Dimensió d_tipus_mobilitat

d_coneixement							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
codi	smallint	2	X			X	
estudi	varchar	57			X	X	

Taula 8. Dimensió d_coneixement

d_entitat							
Atribut	Tipus	Mida	PK	FK	U	NN	AI
rowid	smallint	2	X			X	X
codi	varchar	13				X	
descripcio	varchar	119			X	X	
pais	varchar	2				X	

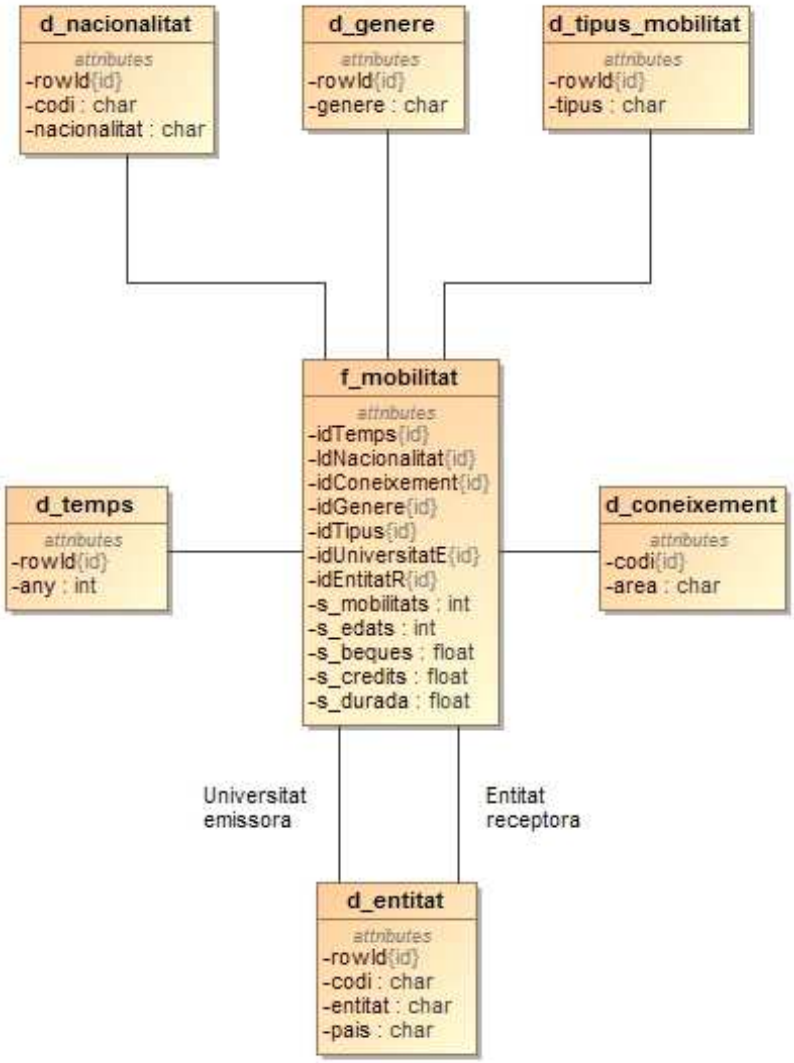
Taula 9. Dimensió d_entitat

Llegenda:

Tipus	Descripció
PK	Clau primaria
FK	Clau forana
U	Únic
NN	No nul
AI	Camp autoincremental

Taula 10. Llegenda

A la il·lustració 6 es pot veure el model físic complet:

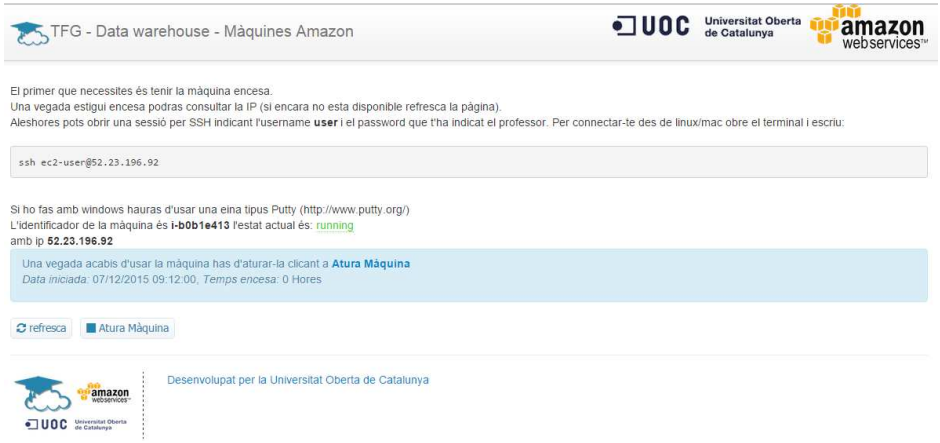


Il·lustració 7. Model final

3.Implementació

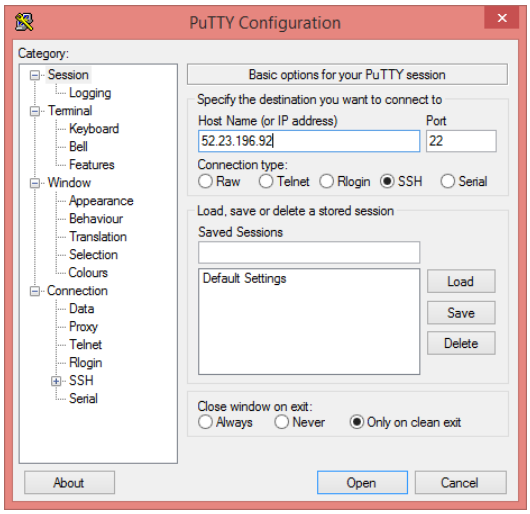
Com s'ha vist al punt 1.4 el programari que s'utilitzarà per implementar el DW està allotjat a Amazon Elastic Compute Cloud (Amazon EC2) que forma part de la plataforma de còmput en el núvol de l'empresa Amazon.com anomenada Amazon Web Services:

- A la il·lustració 8 es veu l'enllaç habilitat a l'aula del TFG per encendre la maquina i obtenir la IP:



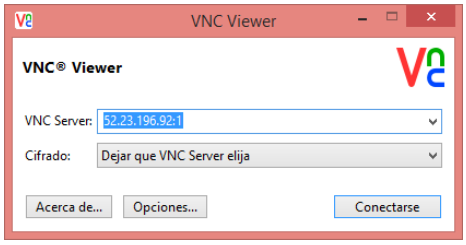
Il·lustració 8. Màquines Amazon

- Amb la IP de la maquina s'obre una sessió SSH amb el programa PuTTY (USER:user, PWD: Student2015*) i s'executa vncserver, com es mostra a la il·lustració 9:

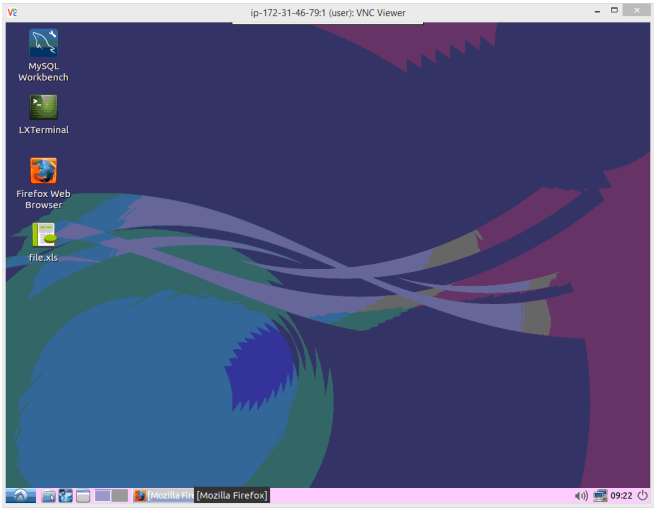


Il·lustració 9. Putty

- Les il·lustracions 10 i 11 mostren el programari VNC Viewer on s'introdueix la IP i el password "Student" per poder visualitzar l'escriptori virtual:



Il·lustració 10. VNC Viewer

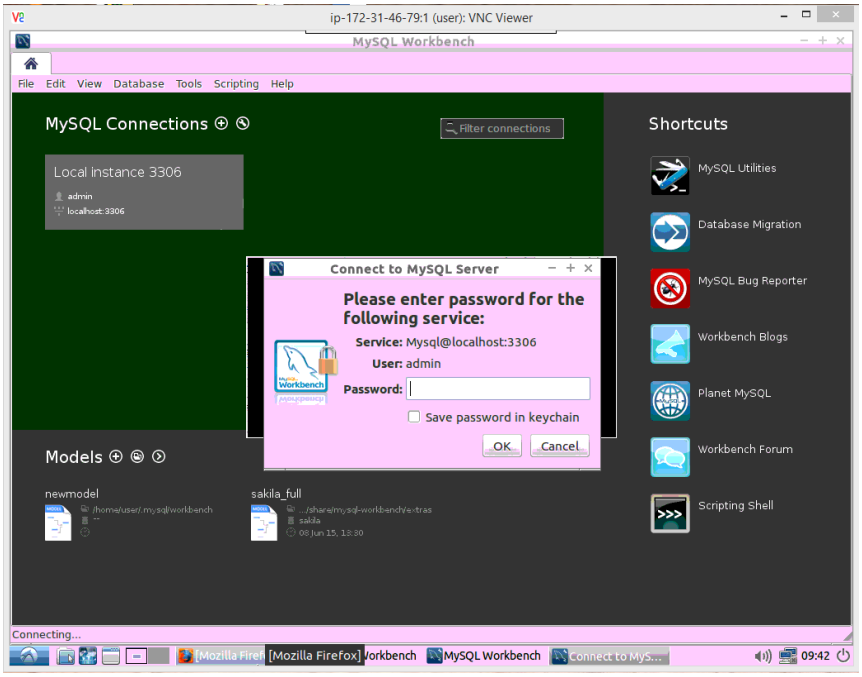


Il·lustració 11. Escriptori virtual

3.1 Creació de la BBDD

La primera acció que es duu a terme és crear la base de dades on s'allotjarà el magatzem de dades amb el programari MySQL Workbench 6.3.4.0.

La il·lustració 12 mostra la connexió amb el servidor MySQL mitjançant el password (dw2015*):



Il·lustració 12. MySQL Workbench

La il·lustració 13 mostra el codi SQL utilitzat per crear la BBDD i l'usuari pel procés ETL amb els seus permisos:

```
create database tfg_dw DEFAULT CHARACTER SET utf8;

create USER 'etl'@'localhost' identified BY 'TFG2015';
GRANT ALL ON tfg_dw.* to 'etl'@'localhost';
```

Il·lustració 13. Codi de creació BBDD

3.2 ETL

El procés ETL s'implementa amb el component PDI Spoon (Kettle 5.4.0.1) de Pentaho (home/user/data-integration/spoon.sh).

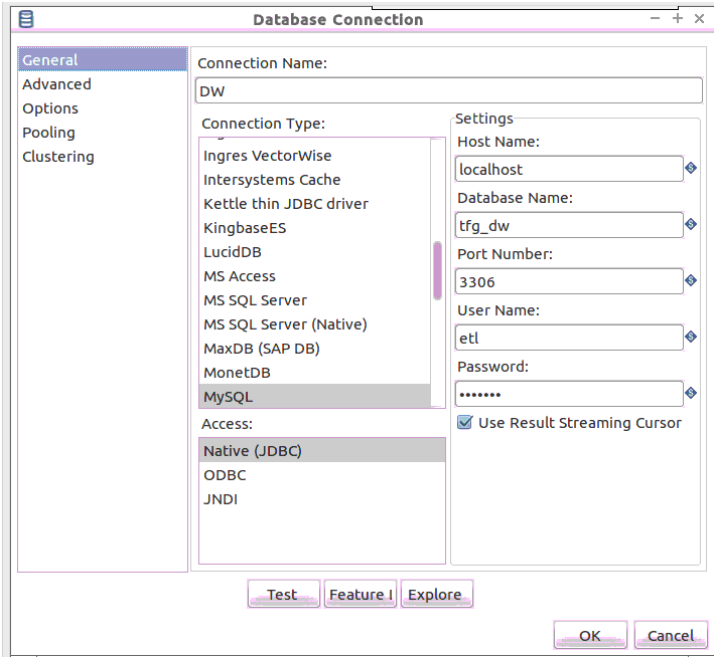
Un dels requisits del treball és que el DW sigui actualitzable automàticament, amb els dos fitxers dels que es disposa es simularà aquest fet. El procés ETL tindrà dues parts: una part de càrrega que només s'executarà una vegada, i la part d'actualització que s'executarà a mesura que es tinguin fitxers de nous cursos.

Job de càrrega: Fitxer SM_2012_13.csv

Job d'actualització:

- Fitxer SM_2012_13.csv i posteriors. Es fa una copia del fitxer d'actualització i es renombra a ACTUALITZACIO_(ANY).csv on (ANY) contindrà el curs de carrega (el fitxer SM_2012_13.csv serà renombrat a ACTUALITZACIO_2012.csv).
- Aquest fitxer es guarda a /home/user/ACTUALITZACIO ja que el procés carrega des d'aquesta carpeta.
- Un cop executat el procés de càrrega s'elimina el fitxer d'actualització o es guarda en una ubicació diferent.
- Els fitxer posteriors a SM_2012_13.csv han de tenir el mateix format que aquest últim (nom de camps, nombre de camps, tipus de dada).

Durant el procés les transformacions necessitaran connexió amb la base de dades que s'ha creat al pas anterior, a la il·lustració 14 es mostra la seva configuració:



Il·lustració 14. Connexió BBDD procés ETL

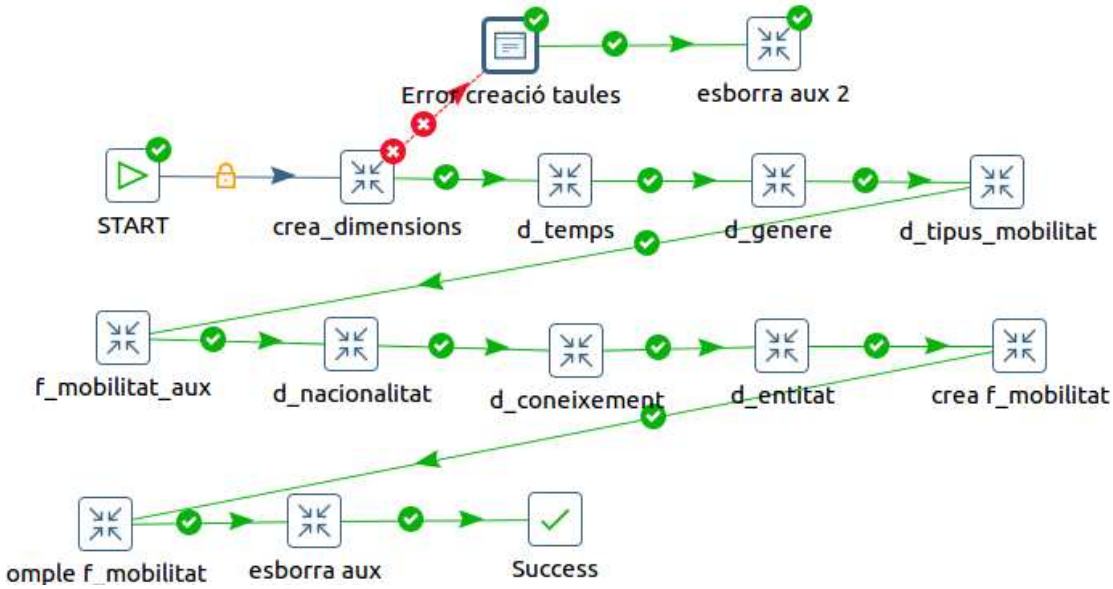
Els jobs i les transformacions es guarden a la carpeta home/user/ETL.

3.2.1 Procés de càrrega:

És l'encarregat d'inserir les dades al DW per primera vegada. La idea general del procés és la de carregar en una taula auxiliar tots els camps del fitxer d'entrada (SM_2012_13.csv) que es necessiten per la construcció de dimensions i fets i a partir d'aquí generar tots els components del DW.

3.2.1.1 Job càrrega:

Conté totes les transformacions per a fer la càrrega. El seu temps d'execució és menor de 2 minuts. La il·lustració 15 mostra el job de càrrega amb el flux de transformacions:

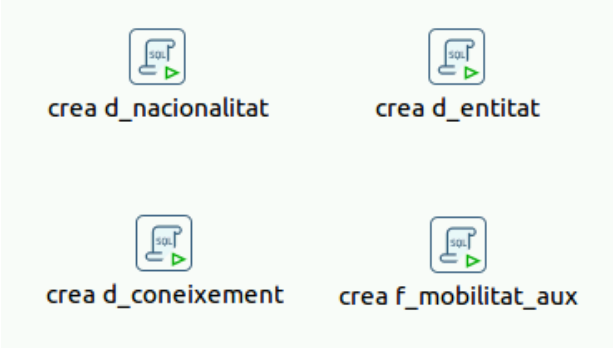


Il·lustració 15. Job càrrega

3.2.1.2 Transformacions:

3.2.1.2.1 crea_dimensions_i_fet

La següent il·lustració mostra la transformació que s'encarrega de crear les taules SQL que requeriran d'una transformació complexa:



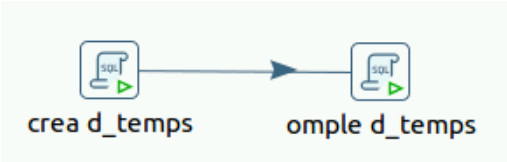
Il·lustració 16. Transformació crea_dimensions_i_fet

Codis SQL de creació de les taules:

```
create table d_nacionalitat(
rowId tinyint not null auto_increment,
codi VARCHAR(2)  not null unique,
descripcio VARCHAR(15)  not null DEFAULT 'ACTUALITZACIO',
constraint pk_d_nacionalitat primary key (rowId));

create table d_entitat(
rowId mediumint not null auto_increment,
codi VARCHAR(300)  not null,
pais VARCHAR(4)  not null,
descripcio VARCHAR(250)  not null DEFAULT 'ACTUALITZACIO' ,
tipus VARCHAR(1) not null DEFAULT 'U',
constraint pk_d_entitat primary key (rowId));

create table d_coneixement(
codi SMALLINT(2)  not null unique,
descripcio VARCHAR(57)  not null DEFAULT 'ACTUALITZACIO' ,
constraint pk_d_coneixement primary key (codi));

create table f_mobilitat_aux(
CURS smallint NOT NULL,
NATIONALITY VARCHAR(2)  NOT NULL,
SUBJECTAREA SMALLINT NOT NULL,
GENDER VARCHAR(1) NOT NULL,
MOBILITYTYPE VARCHAR(1) NOT NULL,
HOMEINSTITUTION VARCHAR(15) NOT NULL,
COUNTRYCODEOFHOMEINSTITUTION VARCHAR(4) NOT NULL,
ENTITATR VARCHAR(300) NOT NULL,
PAISR VARCHAR(2) NOT NULL ,
MOBILITAT SMALLINT NOT NULL,
AGE SMALLINT NOT NULL ,
BEQUES FLOAT  NOT NULL,
TOTALECTSCREDITS FLOAT  NOT NULL,
DURADA FLOAT NOT NULL,
PLACEMENTENTERPRISE VARCHAR(300));
```

3.2.1.2.2 d_temps

La següent il·lustració mostra la transformació que consta de dos passos. Un per la creació de la taula SQL i l'altre que insereix el valor 2011, que també es crea a f_mobilitat_aux.



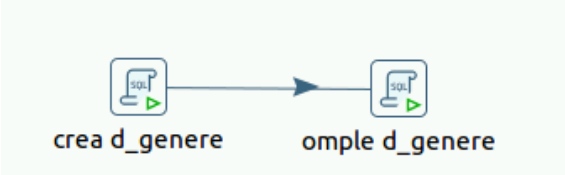
Il·lustració 17. Transformació d_temps

Codi SQL de creació i inserció de la taula:

```
create table d_temps(
rowId tinyint not null auto_increment,
curs smallint not null unique,
constraint pk_d_temps primary key (rowId));

insert into d_temps (curs) VALUE (2011)
```

3.2.1.2.3 d_genere

La següent il·lustració mostra la transformació que consta de dos passos. Un per la creació de la taula SQL i l'altre que insereix els valors 'M' i 'F' i les descripcions. Els valors són només aquests ja que es va comprovar en l'anàlisi de dades de l'entrega anterior.



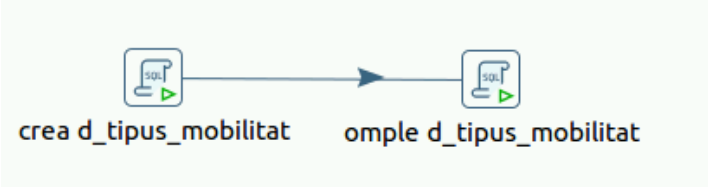
Il·lustració 18. Transformació d_genere

Codi SQL de creació i inserció de la taula:

```
create table d_genere(
rowId TINYINT not null auto_increment,
genere varchar(1) not null unique,
descripcio varchar(13) not null unique DEFAULT 'ACTUALITZACIO',
constraint pk_d_genere primary key (rowId));

insert into d_genere (genere,descripcio) VALUE ('M','MASCULI'),('F','FEMENI');
```

3.2.1.2.4 d_tipus_mobilitat

La il·lustració 19 mostra la transformació que consta de dos passos. Un per la creació de la taula SQL i l'altre que insereix els valors 'S', 'C' i 'P' i les descripcions. Els valors són només aquests ja que es va comprovar en l'anàlisi de dades de l'entrega anterior.



Il·lustració 19. Transformació d_tipus_mobilitat

Codi SQL de creació i inserció de la taula:

```
create table d_tipus_mobilitat(
rowId TINYINT not null auto_increment,
tipus varchar(1) not null unique,
descripcio varchar(13) not null DEFAULT 'ACTUALITZACIO',
constraint pk_d_tipus_mobilitat primary key (rowId));

insert into d_tipus_mobilitat(tipus,descripcio) VALUE
('S','ESTUDI'),('C','MIXTA'),('P','PRACTIQUES');
```

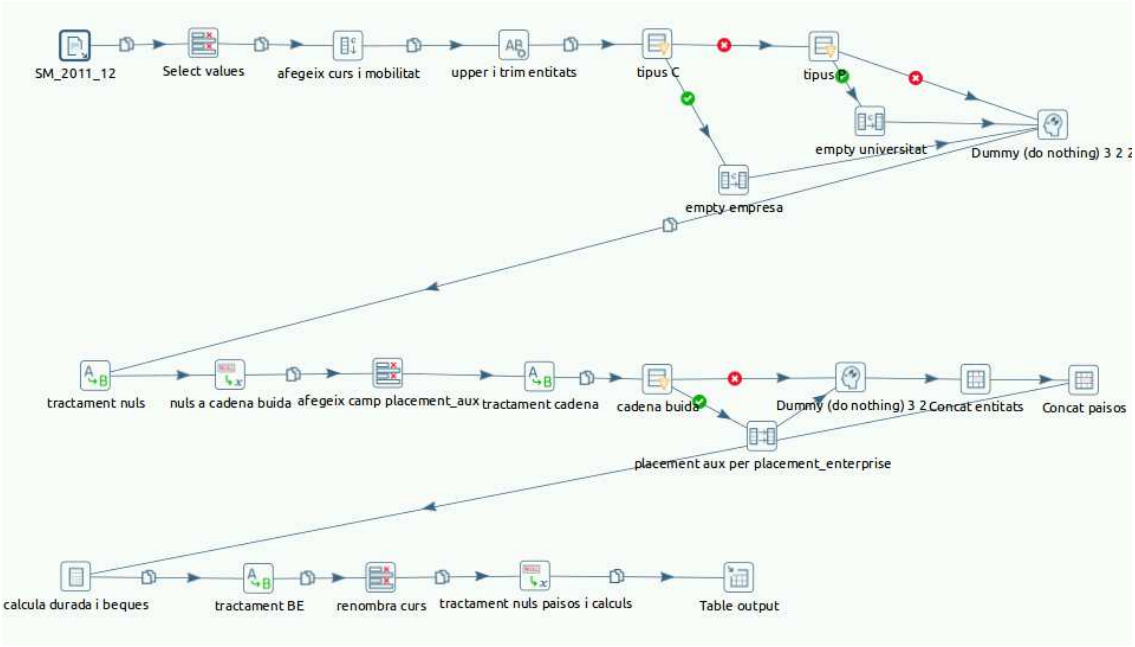
3.2.1.2.5 f_mobilitat_aux

La transformació crea la taula auxiliar que servirà per construir les dimensions i el fet.

Consta de diferents operacions:

- S'importen les dades del fitxer de mobilitats SM_2011_12.csv.
- Es seleccionen els camp necessaris per la construcció de la taula de fet.
- S'afegeixen els camps de curs (2011) i mobilitat (constant a 1).
- Es tracten les cadenes passant a majúscules i eliminant espais davant i darrere.
- En els casos que tinguin els camps d'universitat receptora i empresa informats:
    - Si és una mobilitat de tipus 'C' es seleccionarà la informació de la universitat ja que la d'empresa és opcional.
    - En cas que sigui mobilitat de tipus 'P' es seleccionarà les dades de l'empresa.
- Es tracten els camps que hagin quedat nuls i buits.
- S'aplica el tractament de cadena per tal de crear un codi identificador pels casos en que una empresa tingui descripcions diferents.
- Es concatenen les entitats receptores i els països receptors.
- Es calculen la durada i les beques.
- Es tracten els països BE nombrats com BEFR,BEDE,BENL.
- Tractament final de nuls en països i càlculs.
- Finalment s'insereixen les dades a la taula f_mobilitat_aux.

La il·lustració 20 mostra la transformació completa:



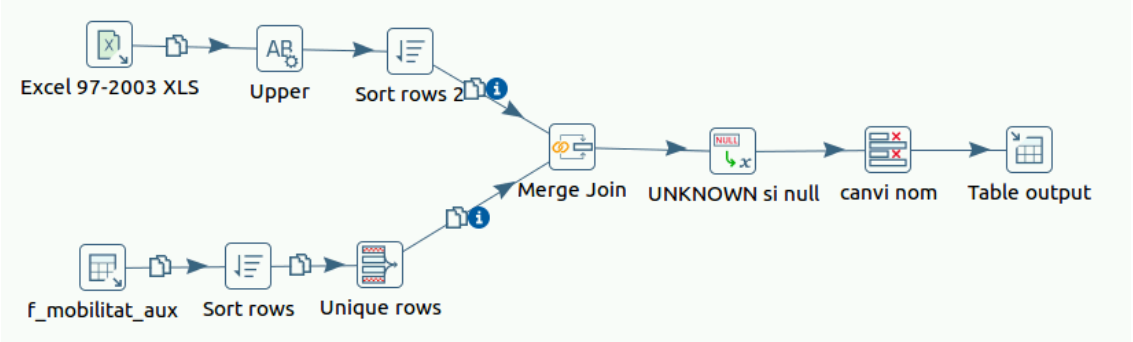
Il·lustració 20. Transformació f_mobilitat_aux

3.2.1.2.6 d_nacionalitat

Per tal de generar aquesta dimensió es necessita importar les dades del fitxer ISOCountryCodes081507.xls i les dades de nacionalitat de la taula f_mobilitat_aux. Seguidament s'executen els següents passos:

- Es passen els camps importats de l'Excel a majúscules.
- Mitjançant una join entre les dues taules es recuperen els camps de descripció per les nacionalitats que es troben en el fitxer f_mobilitat_aux.
- Pels casos en que no es recupera descripció s'assigna a aquest camp el valor 'UNKNOWN'.
- Finalment s'insereixen les dades a la taula d_nacionalitat.

La il·lustració 21 mostra la transformació completa:

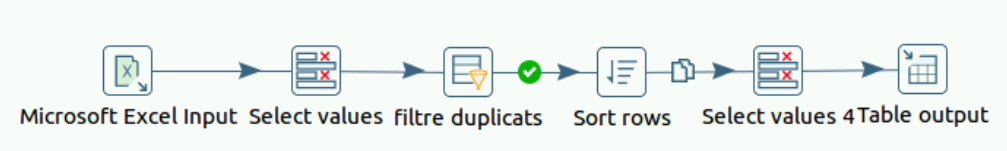


Il·lustració 21. d_nacionalitat

3.2.1.2.7 d_coneixement

Per tal de generar aquesta dimensió es necessita importar les dades del fitxer ISCED97_Erasmus_subject_codes.xls. Es corregeixen els casos susceptibles de crear duplicats detectats a l'entrega anterior. S'insereixen les dades a la taula d_coneixement. No hi ha codis nous al fitxer del curs 2011 per tant no es fa cap comprovació.

La següent il·lustració mostra la transformació completa:



Il·lustració 22. Transformació d_coneixement

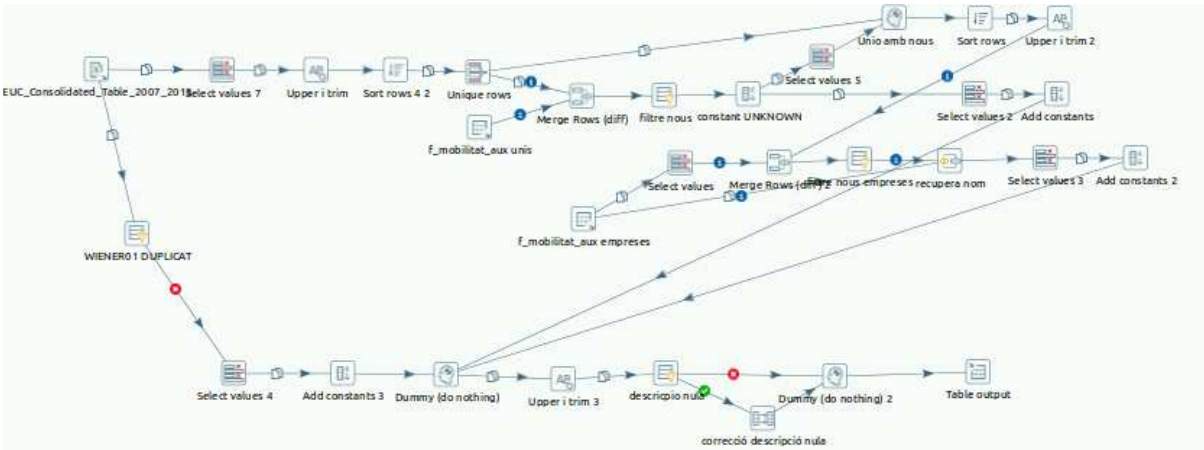
3.2.1.2.8 d_entitat

Per tal de generar aquesta dimensió es necessiten importar les dades (codi i país) del fitxer EUC_Consolidated_Table_2007_2013.xls i les dades de entitats emissores i receptores de la taula f_mobilitat_aux.

Seguidament s'executen els següents passos:

- S'importen les dades d'entitats emissores de la taula f_mobilitat_aux i es comprova quines són noves respecte les importades del fitxer Excel ja que totes dues són dades d'universitats.
- Aquelles que siguin noves són universitats no conegudes, per tant s'assignarà la descripció 'UNKNOWN'.
- S'importen les dades d'entitats receptores de la taula f_mobilitat_aux i es comprova quines són noves respecte les importades del fitxer Excel i les que s'han assignat a 'UNKNOWN' ja que en alguns casos existeixen codis d'universitats al camp d'empresa.
- S'uneixen les dades importades de l'Excel (corregint un cas detectat en l'entrega anterior per evitar duplicats) amb les dades d'universitats noves assignades a 'UNKNOWN' i amb les d'empreses noves.
- S'afegeix la descripció en els casos que sigui nul·la.
- Finalment s'insereixen les dades a la taula d_entitat.

La següent il·lustració mostra la transformació completa:



Il·lustració 23. Transformació d_entitat

3.2.1.2.9 crea f_mobilitat_aux

La següent il·lustració mostra la transformació que crea la taula de fet:

Il·lustració 24. Transformació crea_f_mobilitat

Codi SQL de creació de la taula:

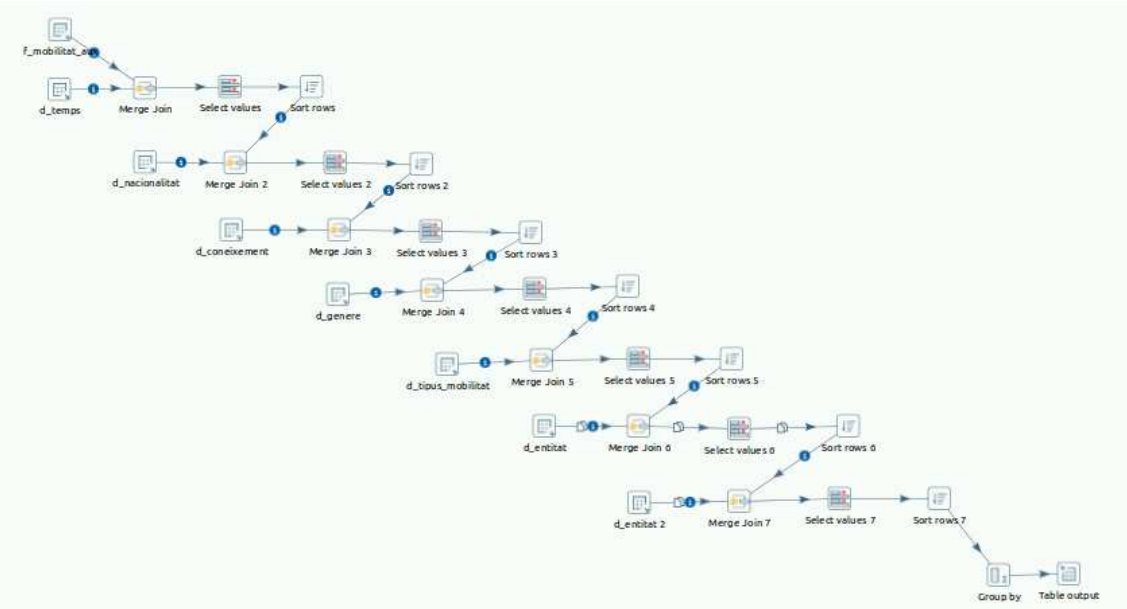
```
CREATE TABLE f_mobilitat(
idTemps TINYINT,
idNacionalitat TINYINT,
idConeixement SMALLINT,
idGenere TINYINT,
idTipus TINYINT,
idUniversitatE MEDIUMINT,
idEntitatR MEDIUMINT,
s_mobilitats SMALLINT,
s_edat SMALLINT,
s_beques FLOAT,
s_credits FLOAT,
s_durada FLOAT,
CONSTRAINT pk_f_mobilitat PRIMARY KEY (idTemps, idNacionalitat,idConeixement,
idGenere, idTipus,idUniversitatE, idEntitatR) ,
CONSTRAINT fk_f_mobilitat_d_temps FOREIGN KEY (idTemps) REFERENCES
d_temps (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_nacionalitat FOREIGN KEY (idNacionalitat)
REFERENCES
d_nacionalitat (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_coneixement FOREIGN KEY (idConeixement)
REFERENCES
d_coneixement (codi) ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_genere FOREIGN KEY (idGenere) REFERENCES
d_genere (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_tipus FOREIGN KEY (idTipus) REFERENCES
d_tipus_mobilitat (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_entitat_E FOREIGN KEY (idUniversitatE)
REFERENCES d_entitat (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT,
CONSTRAINT fk_f_mobilitat_d_entitat_R FOREIGN KEY (idEntitatR) REFERENCES
d_entitat (rowId)
ON DELETE RESTRICT ON UPDATE RESTRICT
);
```

3.2.1.2.10 omple f_mobilitat

Aquesta transformació és l'encarregada de generar les dades per la taula del fet.

S'importen les dades de la taula f_mobilitat_aux i les de totes les dimensions. La filosofia que segueix és fer joins de la taula f_mobilitat_aux amb les taules de dimensió per recuperar els camps de clau primària de les dimensions. Després de fer la ultima join s'obtenen només els camps claus i les mesures. Finalment s'agrupen pels camps claus fent suma de les mesures i s'insereixen les dades a la taula f_mobilitat.

La següent il·lustració mostra la transformació completa:



Il·lustració 25. Transformació omple f_mobilitat

3.2.1.2.11 esborra f_mobilitat_aux

La següent il·lustració mostra la transformació que s'encarrega d'esborrar la taula auxiliar.

Il·lustració 26. Transformació esborra f_mobilitat_aux

Codi SQL d'eliminació de la taula:

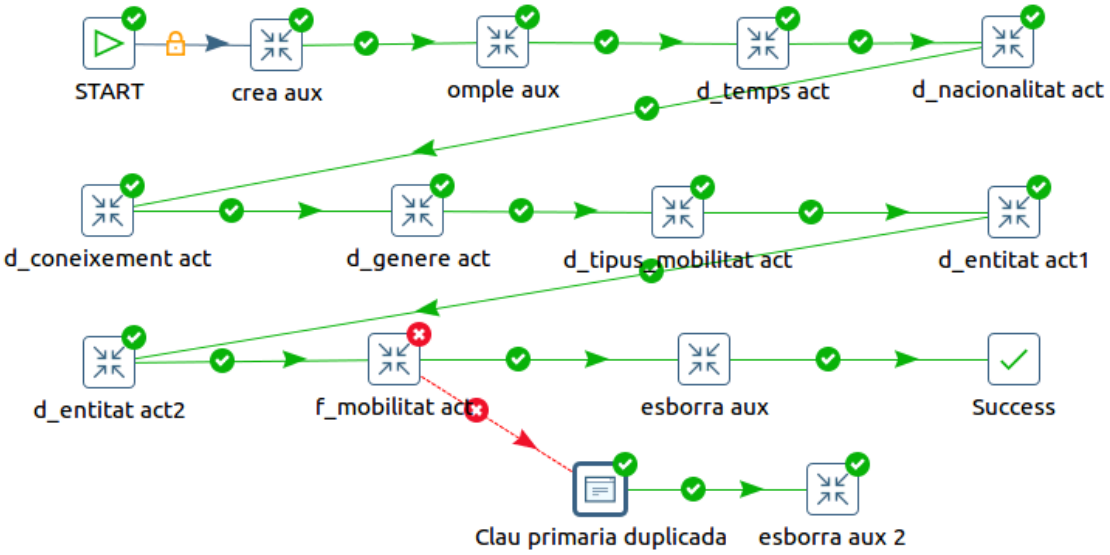
drop table f_mobilitat_aux;

3.2.2 Procés d'actualització:

Es l'encarregat d'inserir al DW els nous cursos. El procés és similar al de càrrega aquest cop comparant les dades de la taula auxiliar amb les de les dimensions ja existents i afegint les dades que siguin noves:

3.2.2.1 Job actualització:

Conté totes les transformacions per a fer l'actualització. El seu temps d'execució és menor de 2 minuts. La il·lustració 27 mostra el job de d'actualització amb el flux de transformacions:

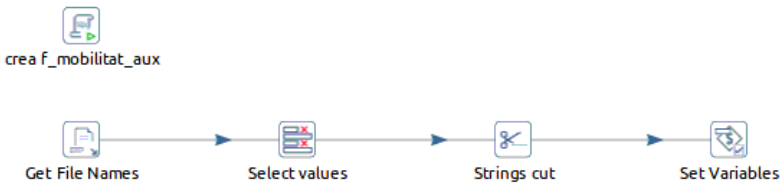


Il·lustració 27. Job actualització

3.2.2.2 Transformacions:

3.2.2.2.1 Crea f_mobilitat_aux_act

La il·lustració 28 mostra la transformació que crea la taula auxiliar. El codi SQL és el mateix que en l'apartat de càrrega.
Addicionalment s'encarrega de recuperar el nom de l'arxiu d'actualització de la carpeta ACTUALITZACIO, extreure'n l'any i guardar-lo en una variable global.



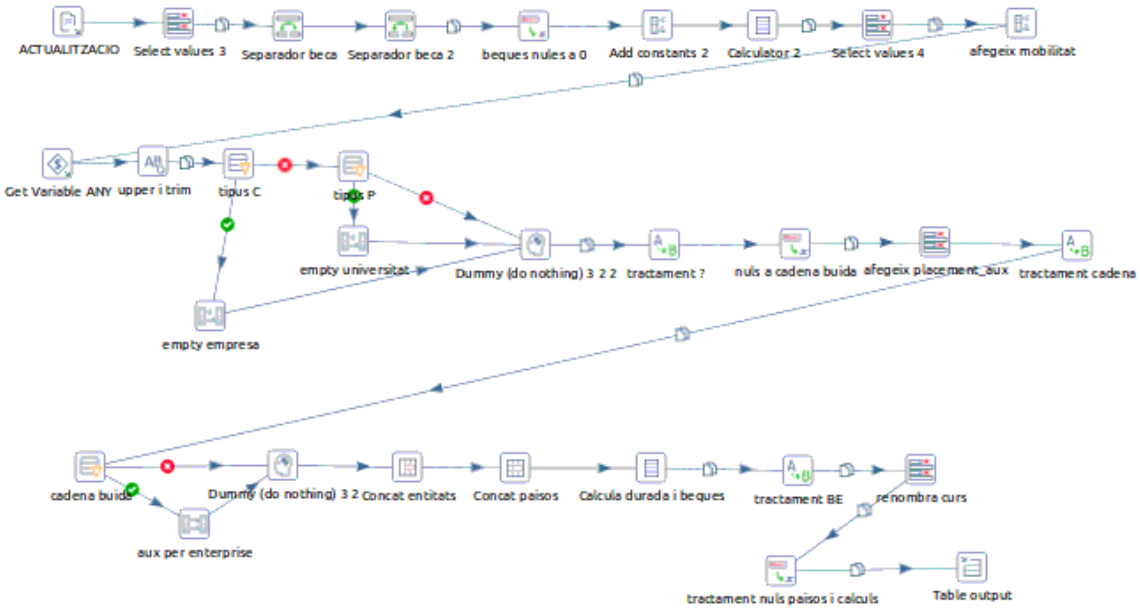
Il·lustració 28. Transformació crea f_mobilitat_aux

3.2.2.2.2 Omple f_mobilitat_aux_act

Transformació que insereix les dades a la taula auxiliar que servirà per construir les dimensions i el fet.

Es seleccionaran els camps que es necessitin del fitxer que es trobi a la carpeta ACTUALITZACIO, es fan les transformacions de cadena necessàries als camps que ho necessitin i es recupera la variable global creada al pas anterior per generar l'any. Respecte al fitxer del 2011 s'han hagut de seleccionar els camps corresponents ja que tenen un nom diferent i fer un tractament pels imports ja que usen un separador decimal no reconegut pel programari.

La següent il·lustració mostra la transformació completa:



Il·lustració 29. Transformació omple f_mobilitat_aux_act

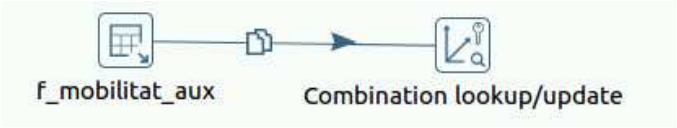
3.2.2.2.3 d_temps_act

La il·lustració 30 mostra la transformació que consta de dos passos. La importació del curs del fitxer auxiliar i l'actualització de la dimensió.

Il·lustració 30. Transformació omple d_temps_act

3.2.2.2.4 d_nacionalitat_act

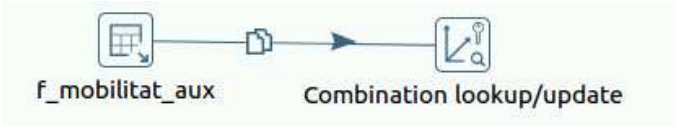
La il·lustració 31 mostra la transformació que consta de dos passos. La importació del codi de nacionalitat del fitxer auxiliar i l'actualització de la dimensió. Aquesta actualització pot ser necessària en el cas que es decidís canviar/afegir codis de nacionalitat o hi hagués un codi erroni, en aquest cas el camp de descripció valdria 'ACTUALITZACIO' tal i com esta definit en la creació de la pròpia taula (DEFAULT VALUE). En aquest cas posteriorment s'hauria de canviar el valor 'ACTUALITZACIÓ' pel que correspongués.



Il·lustració 31. Transformació omple d_nacionalitat_act

3.2.2.2.5 d_coneixement_act

La il·lustració 32 mostra la transformació que consta de dos passos. La importació del codi de coneixement del fitxer auxiliar i l'actualització de la dimensió. Aquesta actualització pot ser necessària en el cas que es decidís canviar els codis de tipus de mobilitat o hi hagués un codi erroni, en aquest cas el camp de descripció valdria 'ACTUALITZACIO' tal i com esta definit en la creació de la pròpia taula (DEFAULT VALUE). En aquest cas posteriorment s'hauria de canviar el valor 'ACTUALITZACIÓ' pel que correspongués.



Il·lustració 32. Transformació d_coneixement_act

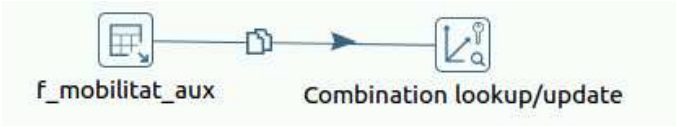
3.2.2.2.6 d_genere_act

La il·lustració 33 mostra la transformació que consta de dos passos. La importació del gènere del fitxer auxiliar i l'actualització de la dimensió. Aquesta actualització pot ser necessària en el cas que es decidís canviar els codis de gènere o hi hagués un codi erroni, en aquest cas el camp de descripció valdria 'ACTUALITZACIO' tal i com esta definit en la creació de la pròpia taula (DEFAULT VALUE). En aquest cas posteriorment s'hauria de canviar el valor 'ACTUALITZACIO' pel que correspongués.

Il·lustració 33. Transformació d_genere_act

3.2.2.2.7 d_tipus_mobilitat_act

La següent il·lustració mostra la transformació que consta de dos passos. La importació del tipus de mobilitat del fitxer auxiliar i l'actualització de la dimensió. Aquesta actualització pot ser necessària en el cas que es decidís canviar els codis de tipus de mobilitat o hi hagués un codi erroni, en aquest cas el camp de descripció valdria 'ACTUALITZACIO' tal i com esta definit en la creació de la pròpia taula (DEFAULT VALUE). En aquest cas posteriorment s'hauria de canviar el valor 'ACTUALITZACIÓ' pel que correspongués.



Il·lustració 34. Transformació omple d_tipus_mobilitat_act

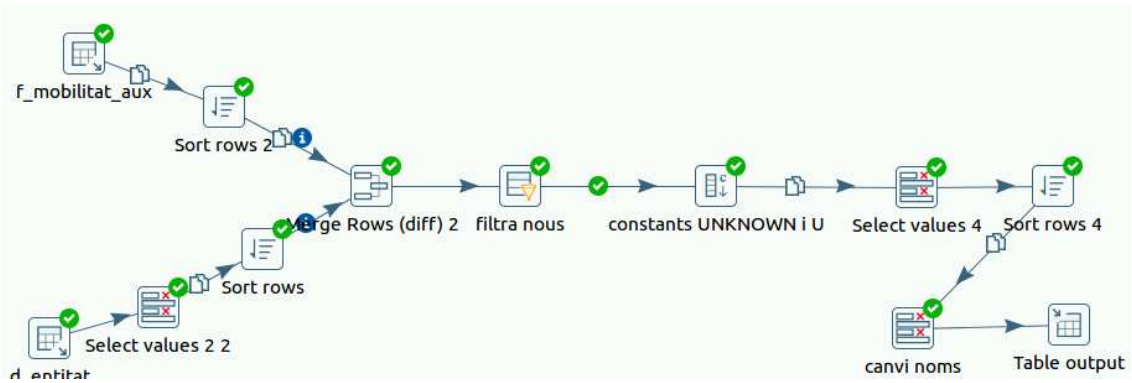
3.2.2.2.8 d_entitat_act

Per tal d'afegir els nous casos a la dimensió d_entiat s'importen les dades (codi i país) de la pròpia dimensió i les dades de entitats emissores de la taula f_mobilitat_aux.

Seguidament es seguiran els següents passos:

- Es comprova quines són noves respecte les existents a la dimensió.
- Aquelles que siguin noves són universitats no conegudes, per tant s'assignaran amb la descripció 'UNKNOWN'
- Finalment s'insereixen les dades noves a la taula d_entitat.

La il·lustració 35 mostra la transformació completa:



Il·lustració 35. Transformació omple d_entitat_act

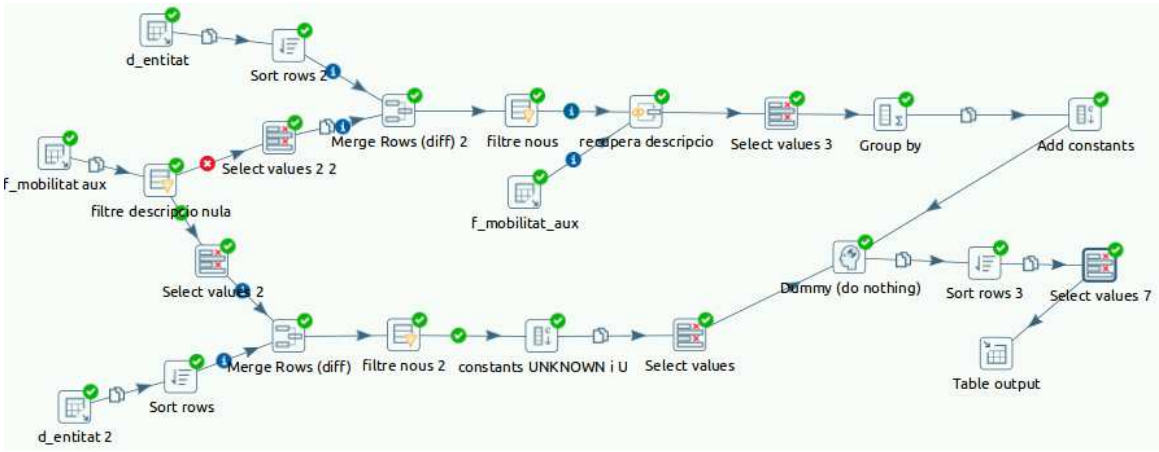
3.2.2.2.9 d_entitat_act2

Per tal d'afegir els nous casos a la dimensió d_entiat s'importen les dades (codi i país) de la pròpia dimensió i les dades de entitats receptores de la taula f_mobilitat_aux.

Seguidament es seguiran els següents passos:

- Es comprova quines són noves respecte les existents a la dimensió.
- Per aquelles que siguin noves es recupera el camp descripció mitjançant una join amb les dades de f_mobilitat_aux.
- Les dades per les que no es recuperi descripció, se'ls i assignarà el codi com a descripció.
- Finalment s'insereixen les dades noves a la taula d_entitat.

La il·lustració 36 mostra la transformació completa:

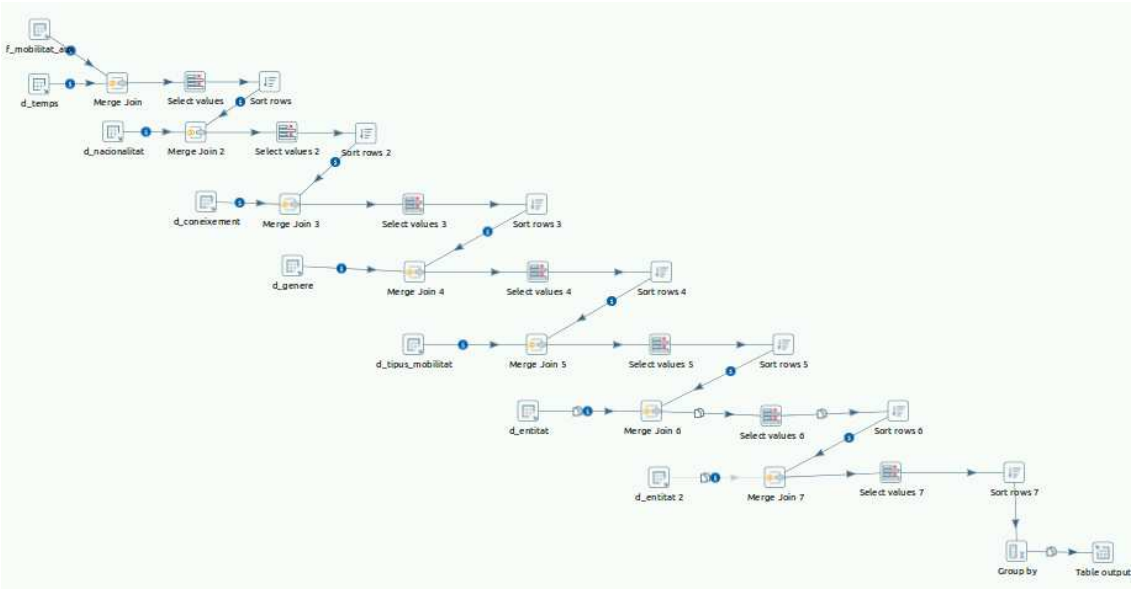


Il·lustració 36. Transformació d_entitat_act2

3.2.2.2.10 f_mobilitat_act

Aquesta transformació és l'encarregada de generar les dades per la taula del fet. Es necessita importar les dades de la taula f_mobilitat_aux i les de totes les dimensions. La filosofia que segueix és fer joins de la taula f_mobilitat_aux amb les taules de dimensió per recuperar els camps de clau primària de les dimensions. Després de fer la ultima joni es tindrà només els camps claus i les mesures. Finalment s'agrupa pels camps claus fent suma de les mesures i s'inserixen les dades a la taula f_mobilitat.

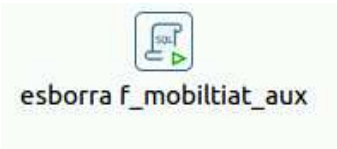
La següent il·lustració mostra la transformació completa:



Il·lustració 37. Transformació f_mobilitat_act

3.2.2.2.11 esborra f_mobilitat_aux

La següent il·lustració mostra la transformació que s'encarrega d'esborrar la taula auxiliar.



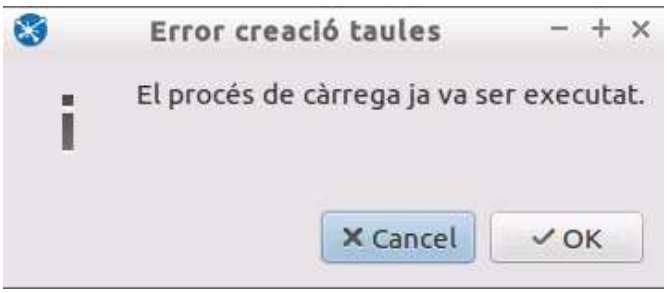
Il·lustració 38. Transformació esborra f_mobilitat_aux

3.2.3 Control d'errors

Tant el procés de carrega com el d'actualització compten amb un control per tal que només puguin ser executats una vegada i no provoquin inconsistències al DW.

El procés de càrrega s'encarrega de crear les taules de dimensió per primera vegada, al tornar-lo a llençar les torna a crear cosa que genera un error en el procés i es mostra el següent missatge:

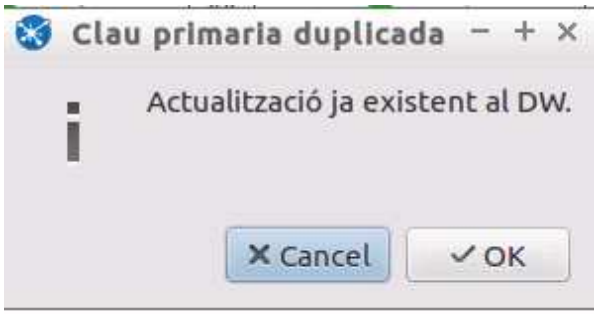
Il·lustració 39 on es mostra l'error de càrrega:



Il·lustració 39. Error de càrrega

Per altra banda al procés d'actualització detecta que torna a ser llançat quan s'intenta inserir les dades a la taula de fet i retorna error perquè la clau primària ja existeix. Les taules de dimensió no es veuran modificades ja que s'actualitzen comprovant les diferencies. Es mostra el següent missatge:

Il·lustració 40 on es mostra l'error d'actualització:



Il·lustració 40. Error d'actualització

Posteriorment als dos missatges s'esborra la taula auxiliar per tal de deixar el procés llest per una pròxima execució.

Addicionalment es pot cometre un error al posar l'any quan es renombra el fitxer d'actualització. Si es carrega el mateix fitxer amb dos anys diferents el procés no ho detecta com a error, però al veure, per exemple, l'informe d'evolució comparativa de mobilitats es tindran dos cursos amb les mateixes mobilitats. En aquests casos caldrà eliminar la BBDD i tornar a generar la creació, càrrega i actualitzacions.

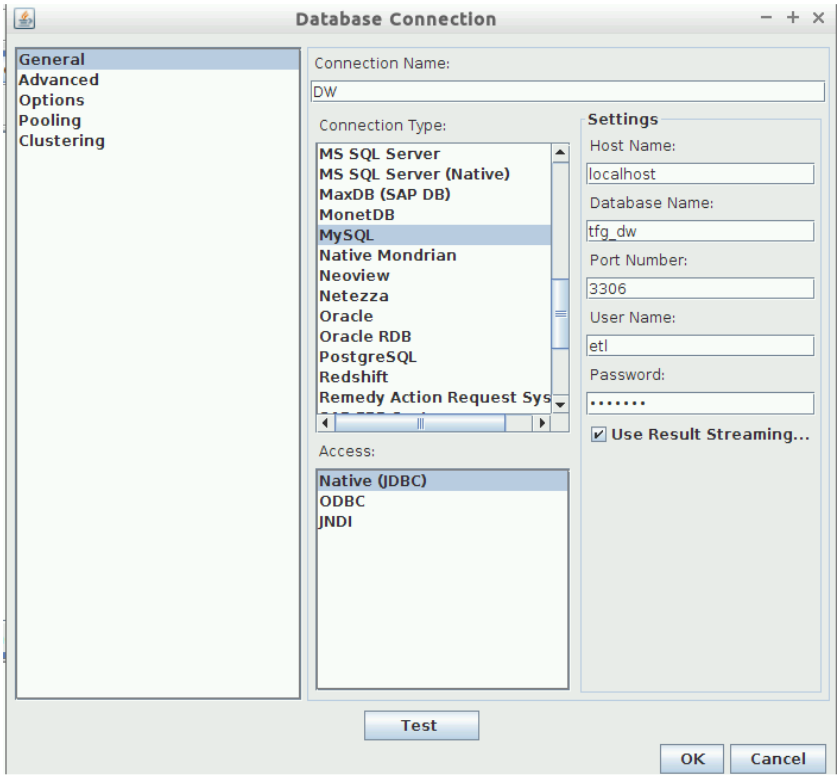
3.3 Cubs OLAP

Un cop carregades les dades es necessari construir els cubs OLAP que ens permetran amb l'eina Saiku fer l'anàlisi i els informes. Saiku és un complement del servidor de BI Pentaho.

Els cubs OLAP es defineixen mitjançant un esquema Mondrian. Mondrian és un servidor OLAP escrit en Java. Permet interactuar amb grans quantitats de dades emmagatzemades en bases de dades relacionals, sense necessitat d'utilitzar sentències SQL .

L'esquema es crea amb el programari Schema Workbench (home/user/schema-workbench/workbench.sh):

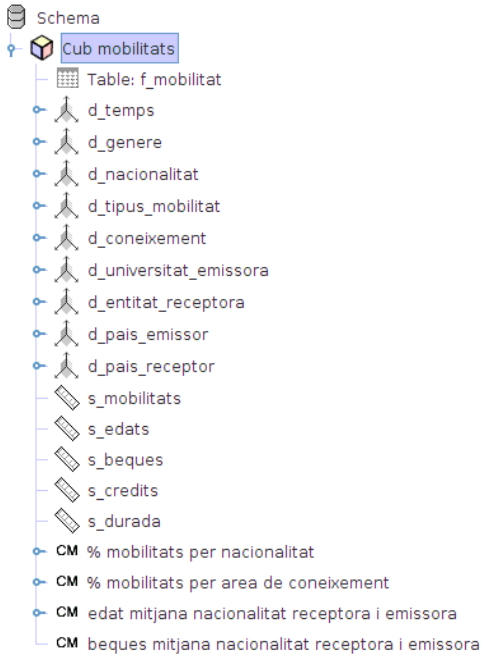
- A la següent il·lustració es mostra la configuració de la connexió amb MySQL per tal d'accedir a les dades per crear l'esquema:



Il·lustració 41. Connexió BBDD Schema Workbench

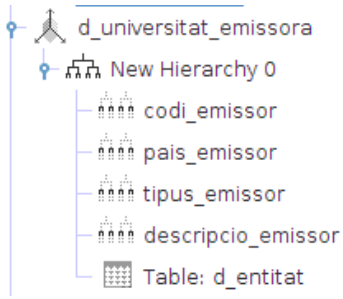
- L'esquema contindrà un cub amb la taula de fets, les dimensions i les mesures que ja s'han creat. A més s'hauran de crear algunes dimensions auxiliars i mesures calculades per tal de donar resposta als informes que calculen percentatges o mitjanes.

  La il·lustració 42 mostra l'esquema final:



Il·lustració 42. Esquema Mobilitats

- La il·lustració 43 mostra els elements que componen una de les dimensions:



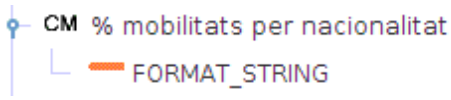
Il·lustració 43. Exemple dimensió d'un esquema

- La il·lustració 44 mostra el codi XML d'una dimensió:

```
−<Dimension type="StandardDimension" visible="true" foreignKey="idUniversitatE" highCardinality="false"
 name="d_universitat_emissora">
 −<Hierarchy name="New Hierarchy 0" visible="true" hasAll="true" primaryKey="rowId">
    <Table name="d_entitat"> </Table>
    <Level name="codi_emissor" visible="true" column="codi" type="String" uniqueMembers="false"
    levelType="Regular" hideMemberIf="Never"> </Level>
    <Level name="pais_emissor" visible="true" column="pais" type="String" uniqueMembers="false"
    levelType="Regular" hideMemberIf="Never"> </Level>
    <Level name="tipus_emissor" visible="true" column="tipus" type="String" uniqueMembers="false"
    levelType="Regular" hideMemberIf="Never"> </Level>
    <Level name="descripcio_emissor" visible="true" column="descripcio" type="String"
    uniqueMembers="false" levelType="Regular" hideMemberIf="Never"> </Level>
  </Hierarchy>
</Dimension>
```

Il·lustració 44. Codi XML de la dimensió d'exemple

- Les següents il·lustracions mostren els elements d'una mesura calculada:



Il·lustració 45. Exemple Calculated member

Calculated Member for 'Cub mobilitats' Cube

Attribute	Value
name	% mobilitats per nacionalitat
description	
caption	
dimension	Measures
hierarchy	
parent	
visible	
formula | formulaElem...	[Measures].[s_mobilitats]/ ([d_nacionalitat.New Hierarchy 0].[All d_nacionalitat.New...
formatString	

Il·lustració 46. Exemple Calculated member 2

- La il·lustració 47 mostra el codi XML d'una mesura calculada:

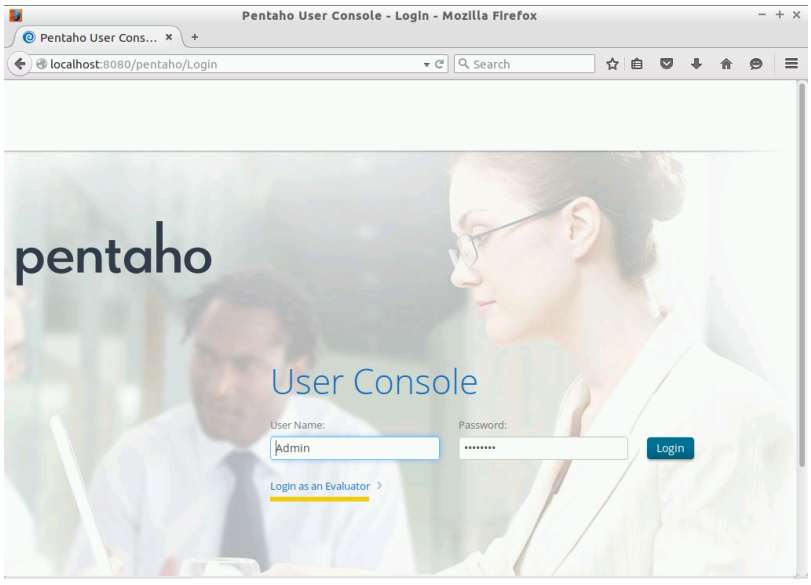
```
−<CalculatedMember name="% mobilitats per nacionalitat" formula="[Measures].[s_mobilitats]/
([d_nacionalitat.New Hierarchy 0].[All d_nacionalitat.New Hierarchy 0s] , [Measures].[s_mobilitats])"
dimension="Measures">
   <CalculatedMemberProperty name="FORMAT_STRING" value="0.0%"> </CalculatedMemberProperty>
</CalculatedMember>
```

Il·lustració 47. Codi XML del calculated membre d'exemple

L'esquema resultant esta ubicat a la carpeta home/ESQUEMES/Schema_Mobilitats.xml.

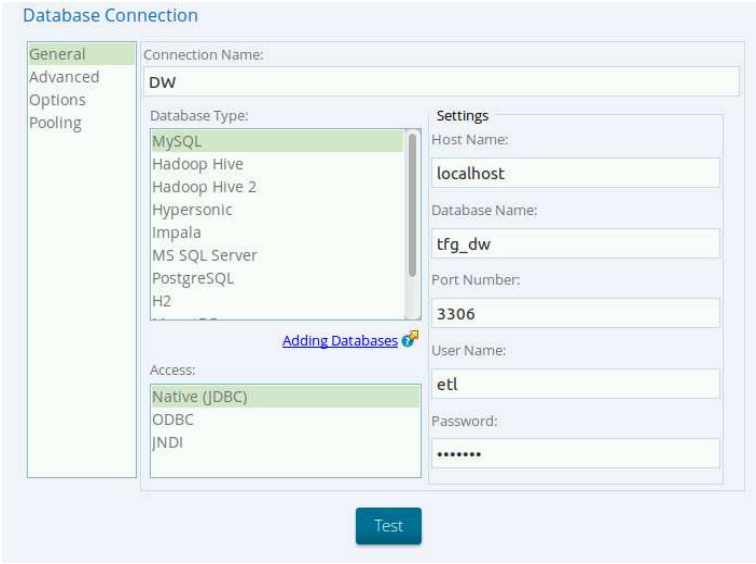
Un cop creat l'esquema, s'ha d'incloure al repositori d'esquemes del servidor Pentaho, però abans és necessari fer la connexió amb el DW ja que no esta allotjat al mateix servidor:

- La il·lustració 48 mostra el navegador web amb el servidor de Pentaho després d'executar home/user/biserver-ce/start-pentaho.sh:



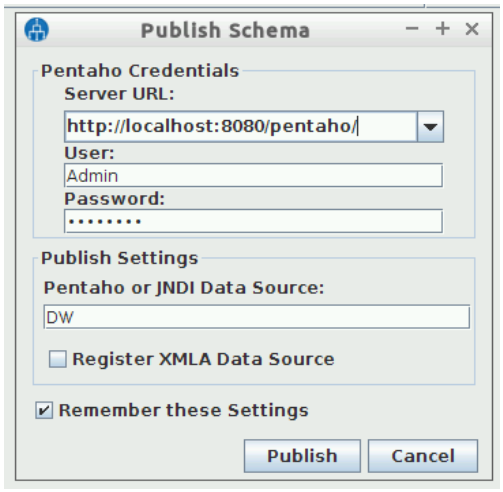
Il·lustració 48. Pentaho

- La il·lustració 49 mostra el menú “manage data sources” on es configura la connexió:



Il·lustració 49. Connexió BBDD Pentaho

- Un cop creada la connexió de Pentaho ja es pot publicar l’esquema al servidor, com es veu a la il·lustració 50, mitjançant l’opció del menú ‘publish’ de Schema Workbench:

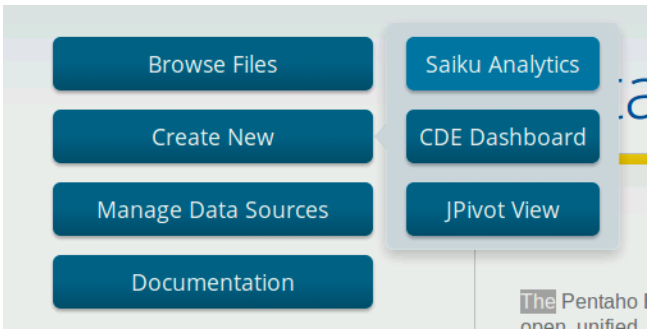


Il·lustració 50. Publicació de l’esquema al servidor Pentaho

3.4 Informes

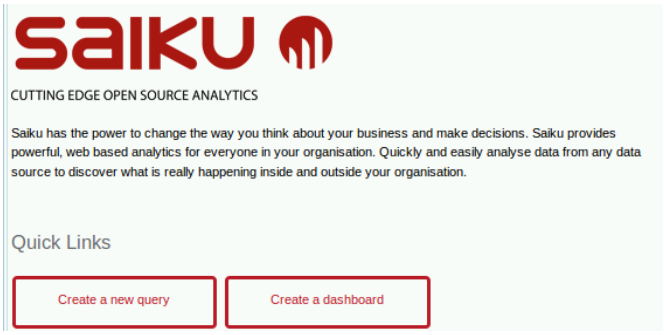
Finalment es torna al servidor Pentaho i es segueixen els següents passos per visualitzar el informes:

- La Il·lustració 51 mostra l'accés a Saiku mitjançant la pàgina principal de Pentaho:



Il·lustració 51. Accés a Saiku des de Pentaho

- La Il·lustració 52 mostra la creació d'una nova query per accedir a l'espai de visualització on es creen els informes:



Il·lustració 52. Query Saiku

Els informes es guarden a la plataforma Pentaho a una carpeta que es troba a public/Informes.

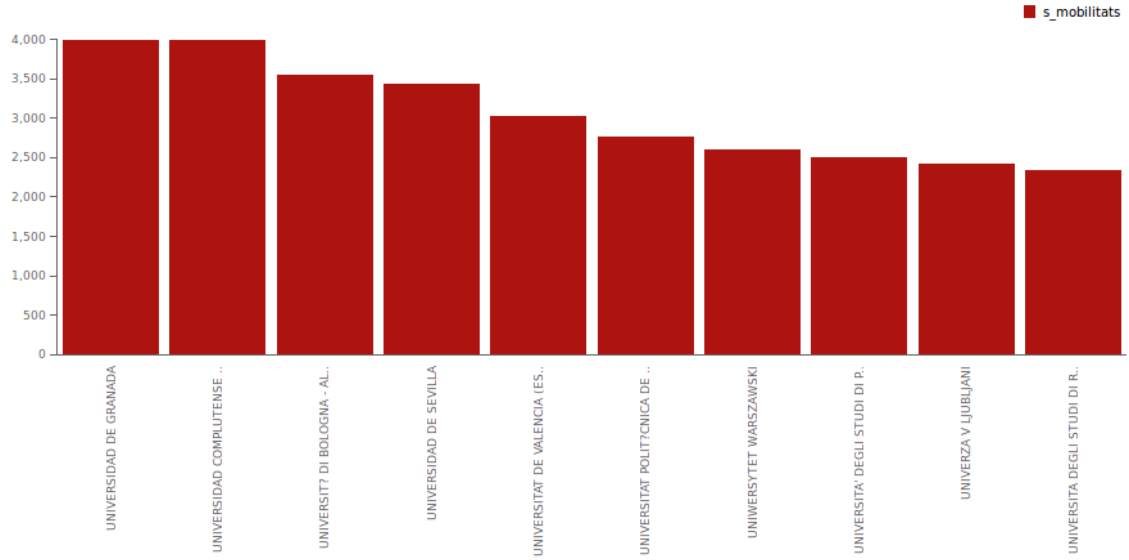
Addicionalment els informes també es troben exportats a Ubuntu en format .xlsx (home/user/Informes).

Per tal de no extendre massa el document només es mostra una vegada el detall del codi MDX d'aquells informes que són semblants. També es mostren alguns exemples de gràfics repartits en els diferents informes.

3.4.1 Top 10 d'universitats més receptores, i més emissores de mobilitats

La il·lustració 53 mostra l'informe d'universitats emissores top 10 sobre el total de dades:

descripcio_emissor	s_mobilitats
UNIVERSIDAD DE GRANADA	3,999
UNIVERSIDAD COMPLUTENSE DE MADRID	3,994
UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	3,543
UNIVERSIDAD DE SEVILLA	3,437
UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	3,035
UNIVERSITAT POLIT?CNICA DE VALENCIA	2,771
UNIWERSYTET WARSZAWSKI	2,597
UNIVERSITA' DEGLI STUDI DI PADOVA	2,502
UNIVERZA V LJUBLJANI	2,430
UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA	2,335



Il·lustració 53. Top 10 emissores total

La il·lustració 54 mostra la query MDX de l'informe anterior:

```
WITH
SET [~ROWS] AS
   TopCount({[d_universitat_emissora.New Hierarchy 0].[descripcio_emissor].Members},
10, [Measures].[s_mobilitats])
SELECT
NON EMPTY {[Measures].[s_mobilitats]} ON COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 54. Query top 10 emissores total

La il·lustració 55 mostra l'informe d'universitats emissores top 10 del curs 2011:

curs	descripcio_emissor	s_mobilitats
2011	UNIVERSIDAD DE GRANADA	2,101
	UNIVERSIDAD COMPLUTENSE DE MADRID	2,065
	UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	1,713
	UNIVERSIDAD DE SEVILLA	1,694
	UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	1,532
	UNIVERSITAT POLIT?CNICA DE VALENCIA	1,466
	UNIWERSYTET WARSZAWSKI	1,349
	UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA	1,213
	UNIVERSITA' DEGLI STUDI DI PADOVA	1,195
	UNIVERZA V LJUBLJANI	1,188

Il·lustració 55. Top 10 emissores 2011

La il·lustració 56 mostra la query MDX de l'informe anterior:

```
WITH
SET [~ROWS_d_temps_d_temps.New Hierarchy 0] AS
   {[d_temps.New Hierarchy 0].[2011]}
SET [~ROWS_d_universitat_emissora_d_universitat_emissora.New Hierarchy 0] AS
   {[d_universitat_emissora.New Hierarchy 0].[descripcio_emissor].Members}
SELECT
NON EMPTY {[Measures].[s_mobilitats]} ON COLUMNS,
NON EMPTY TopCount(NonEmptyCrossJoin([~ROWS_d_temps_d_temps.New Hierarchy
0], [~ROWS_d_universitat_emissora_d_universitat_emissora.New Hierarchy 0]), 10,
[Measures].[s_mobilitats]) ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 56. Query top 10 emissores 2011

La il·lustració 57 mostra l'informe d'universitats emissores top 10 del curs 2012:

curs	descripcio_emissor	s_mobilitats
2012	UNIVERSIDAD COMPLUTENSE DE MADRID	1,929
	UNIVERSIDAD DE GRANADA	1,898
	UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	1,830
	UNIVERSIDAD DE SEVILLA	1,743
	UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	1,503
	UNIVERSITA' DEGLI STUDI DI PADOVA	1,307
	UNIVERSITAT POLIT?CNICA DE VALENCIA	1,305
	UNIWERSYTET WARSZAWSKI	1,248
	UNIVERZA V LJUBLJANI	1,242
	UNIVERSITAET WIEN	1,183

Il·lustració 57. Top 10 emissores 2012

La il·lustració 58 mostra l'informe d'universitats receptores top 10 sobre el total de dades:

descripcio	s_mobilitats
UNIVERSIDAD DE GRANADA	4,005
UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	3,473
UNIVERSIDAD DE SEVILLA	3,459
UNIVERSIDAD COMPLUTENSE DE MADRID	3,359
UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	3,313
UNIVERSITAT POLIT?CNICA DE VALENCIA	2,858
UNIVERZITA KARLOVA V PRAZE	2,453
UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA	2,240
UNIVERSITAT DE BARCELONA	2,201
UNIVERSIDAD DE SALAMANCA	2,146

Il·lustració 58. Top 10 receptores total

La il·lustració 59 mostra l'informe d'universitats receptores top 10 del curs 2011:

curs	descripcio	s_mobilitats
2011	UNIVERSIDAD DE GRANADA	2,052
	UNIVERSIDAD DE SEVILLA	1,769
	UNIVERSIDAD COMPLUTENSE DE MADRID	1,709
	UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	1,698
	UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	1,693
	UNIVERSITAT POLIT?CNICA DE VALENCIA	1,508
	UNIVERZITA KARLOVA V PRAZE	1,137
	UNIVERSIDAD DE SALAMANCA	1,110
	UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA	1,107
	UNIVERSITAT DE BARCELONA	1,105

Il·lustració 59. Top 10 receptores 2011

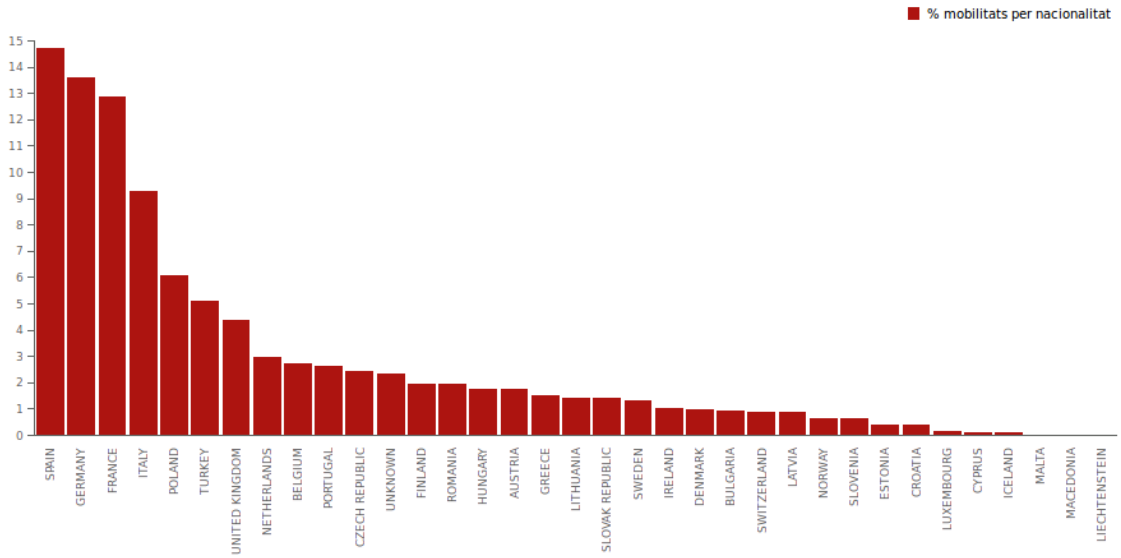
La il·lustració 60 mostra l'informe d'universitats receptores top 10 del curs 2012:

curs	descripcio	s_mobilitats
2012	UNIVERSIDAD DE GRANADA	1,953
	UNIVERSITAT DE VALENCIA (ESTUDI GENERAL) UVEG	1,775
	UNIVERSIDAD DE SEVILLA	1,690
	UNIVERSIDAD COMPLUTENSE DE MADRID	1,650
	UNIVERSIT? DI BOLOGNA - ALMA MATER STUDIORUM	1,620
	UNIVERSITAT POLIT?CNICA DE VALENCIA	1,350
	UNIVERZITA KARLOVA V PRAZE	1,316
	UNIVERSITA DEGLI STUDI DI ROMA LA SAPIENZA	1,133
	UNIVERSITAT DE BARCELONA	1,096
	UNIVERZA V LJUBLJANI	1,074

Il·lustració 60. Top 10 receptores 2012

3.4.2 Distribució en % de mobilitats per nacionalitat

La il·lustració 61 mostra l'informe de distribució de mobilitats per nacionalitat total:



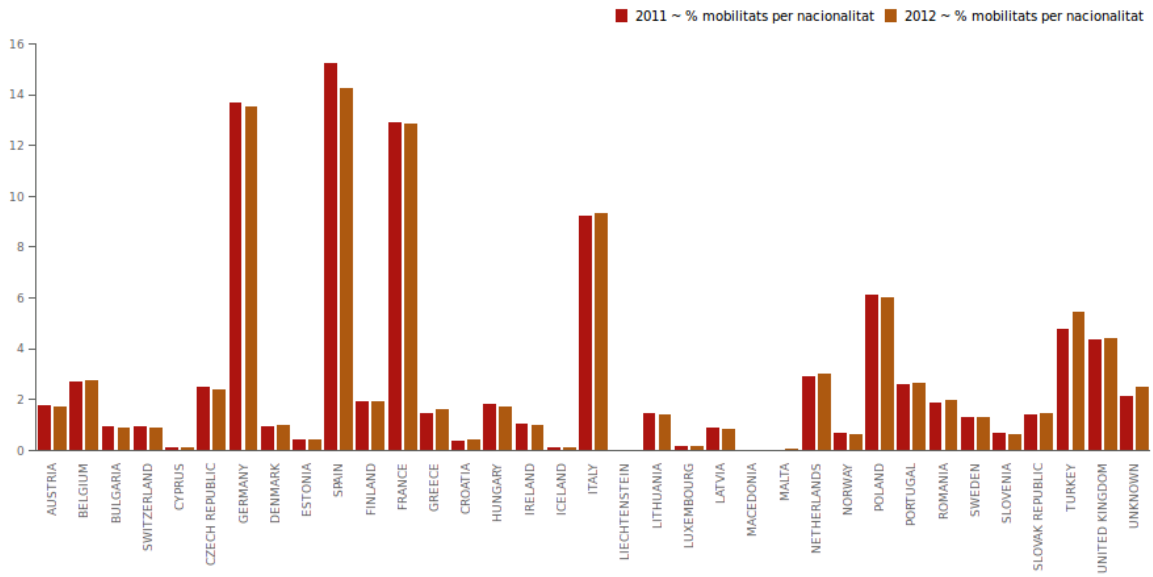
Il·lustració 61. Distribució en % de mobilitats per nacionalitat total

La il·lustració 62 mostra la query MDX de l'informe anterior:

```
WITH
SET [~ROWS] AS
   Order({[d_nacionalitat.New Hierarchy 0].[descripcio].Members}, [Measures].[%
mobilitats per nacionalitat], DESC)
SELECT
NON EMPTY {[Measures].[% mobilitats per nacionalitat]} ON COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 62. Query distribució en % de mobilitats per nacionalitat total

La il·lustració 63 mostra l'informe de distribució de mobilitats per nacionalitat detall per curs:



Il·lustració 63. Distribució en % de mobilitats detall per curs

La il·lustració 64 mostra la query MDX de l'informe anterior:

```
WITH
SET [~COLUMNS] AS
   {[d_temps.New Hierarchy 0].[curs].Members}
SET [~ROWS] AS
   {[d_nacionalitat.New Hierarchy 0].[descripcio].Members}
SELECT
NON EMPTY CrossJoin([~COLUMNS], {[Measures].[% mobilitats per nacionalitat]}) ON
COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 64. Query distribució en % de mobilitats detall per curs

3.4.3 Distribució en % de mobilitats per àrea de coneixement

L'informe és difícil de mostrar ja que les àrees de coneixement són nombroses.

La il·lustració 65 mostra l'informe de distribució de mobilitats per coneixement total:

descripcio	% mobilitats per area de coneixement
Foreign languages	9.0%
Business and administration	8.2%
Business and administration (broad programmes)	7.4%
Law	5.4%
Engineering and engineering trades	4.0%
Economics	3.5%
Humanities	3.2%
Political science and civics	3.1%
Medicine	2.4%
Architecture and town planning	2.4%
Social and behavioural science	2.0%
Management and administration	1.9%
Computer science	1.6%
Architecture and building	1.6%
Mother tongue	1.6%
Biology and biochemistry	1.6%
Mechanics and metal work	1.5%
Design	1.5%
Journalism and information	1.4%
Psychology	1.4%
Travel, tourism and leisure	1.3%
Building and civil engineering	1.2%

Il·lustració 65. Distribució en % de mobilitats per àrea de coneixement total

La il·lustració 66 mostra l'informe de distribució de mobilitats per coneixement detall per curs:

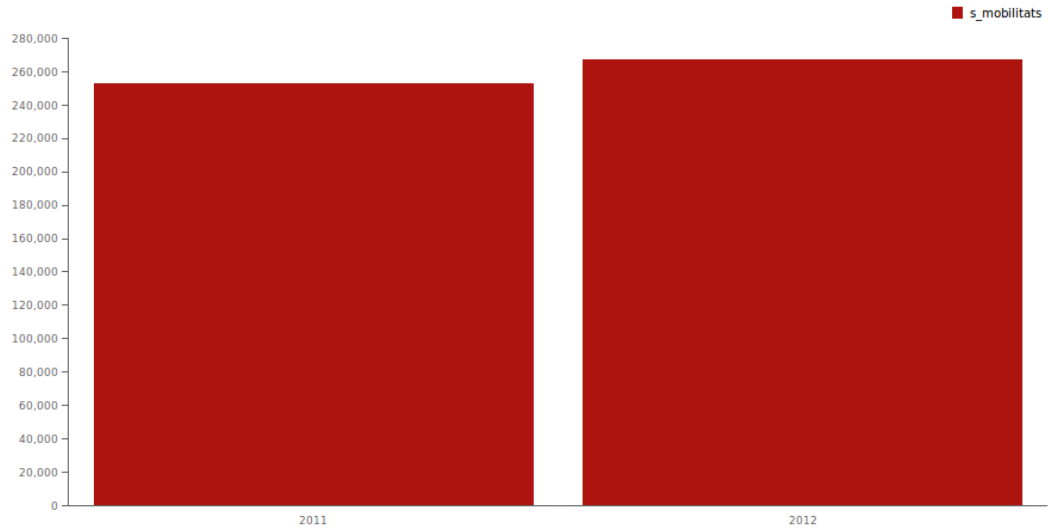
	2011	2012
descripcio	% mobilitats per area de coneixement	% mobilitats per area de coneixement
Foreign languages	9.6%	8.3%
Business and administration	8.7%	7.8%
Business and administration (broad programmes)	7.3%	7.5%
Law	5.5%	5.4%
Engineering and engineering trades	3.9%	4.0%
Economics	3.8%	3.2%
Humanities	2.3%	3.9%
Political science and civics	3.2%	2.9%
Medicine	2.6%	2.3%
Architecture and town planning	2.7%	2.0%
Social and behavioural science	1.8%	2.2%
Management and administration	1.8%	1.9%
Computer science	1.8%	1.5%
Architecture and building	1.2%	2.0%
Mother tongue	1.6%	1.5%
Biology and biochemistry	1.7%	1.5%
Mechanics and metal work	1.6%	1.4%
Design	1.5%	1.5%
Journalism and information	1.5%	1.4%

Il·lustració 66. Distribució en % de mobilitats per àrea detall per curs

3.4.4 Evolució comparativa del nombre de mobilitats per curs

La il·lustració 67 mostra l'informe de comparativa de mobilitats per curs:

curs	s_mobilitats
2011	252,827
2012	267,547



Il·lustració 67. Evolució Comparativa del nombre de mobilitats per curs

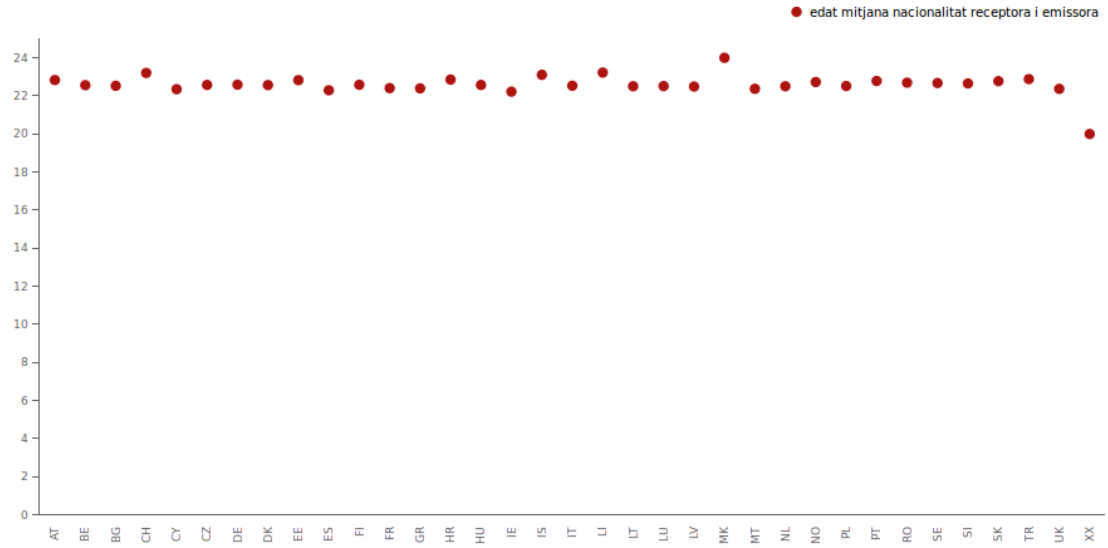
La il·lustració 68 mostra la query MDX de l'informe anterior:

```
WITH
SET [~ROWS] AS
  {[d_temps.New Hierarchy 0].[curs].Members}
SELECT
NON EMPTY {[Measures].[s_mobilitats]} ON COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 68. Query evolució Comparativa del nombre de mobilitats per curs

3.4.5 Edat mitjana per nacionalitat receptora i emissora

La il·lustració 69 mostra l'informe d'edat mitjana per nacionalitat receptora total:



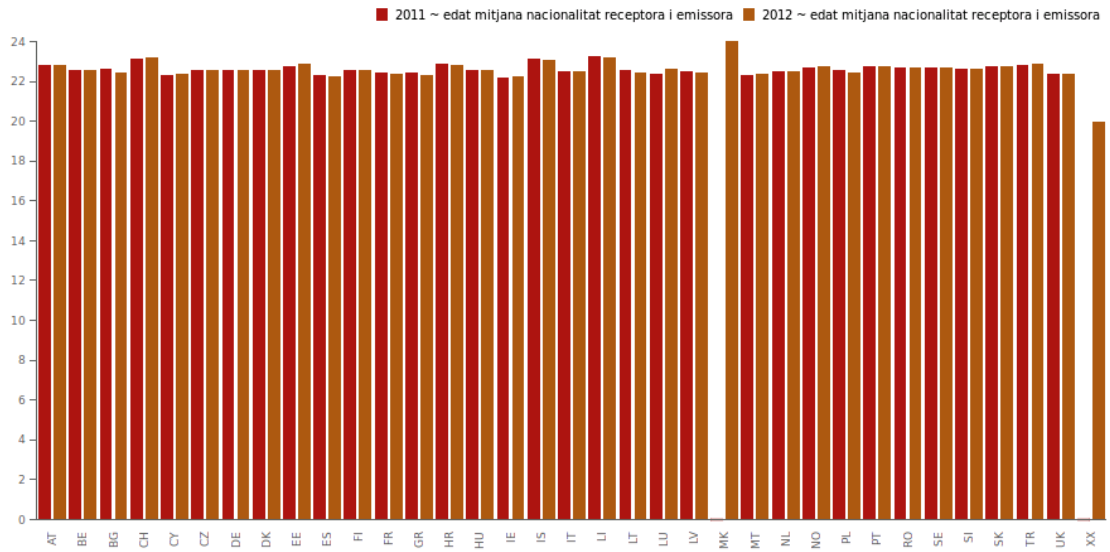
Il·lustració 69. Edat mitjana per nacionalitat receptora total

La il·lustració 70 mostra la query MDX de l'informe anterior:

```
WITH
SET [~ROWS] AS
   {[d_pais_receptor.New Hierarchy 0].[pais receptor].Members}
SELECT
NON EMPTY {[Measures].[edat mitjana nacionalitat receptora i emissora]} ON COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 70. Query edat mitjana per nacionalitat receptora total

La il·lustració 71 mostra l'informe d'edat mitjana per nacionalitat receptora detall per curs:



Il·lustració 71. Edat mitjana per nacionalitat receptora detall

La il·lustració 72 mostra la query MDX de l'informe anterior:

```
WITH
SET [~COLUMNS] AS
   {[d_temps.New Hierarchy 0].[curs].Members}
SET [~ROWS] AS
   {[d_pais_receptor.New Hierarchy 0].[pais receptor].Members}
SELECT
NON EMPTY CrossJoin([~COLUMNS], {[Measures].[edat mitjana nacionalitat receptora i
emissora]}) ON COLUMNS,
NON EMPTY [~ROWS] ON ROWS
FROM [Cub mobilitats]
```

Il·lustració 72. Query edat mitjana per nacionalitat receptora detall

La il·lustració 73 mostra l'informe d'edat mitjana per nacionalitat emissora total:

pais_emissor	edat mitjana nacionalitat receptora i emissora
AT	23.4
BE	21.7
BG	22.7
CH	23.5
CY	21.0
CZ	23.3
DE	23.4
DK	23.9
EE	23.0
ES	22.6
FI	23.7
FR	21.3
GR	22.3
HR	23.3
HU	22.8
IE	21.7
IS	25.4
IT	23.1
LI	25.5
LT	21.5
LU	22.0
LV	22.6

Il·lustració 73. Edat mitjana per nacionalitat emissora total

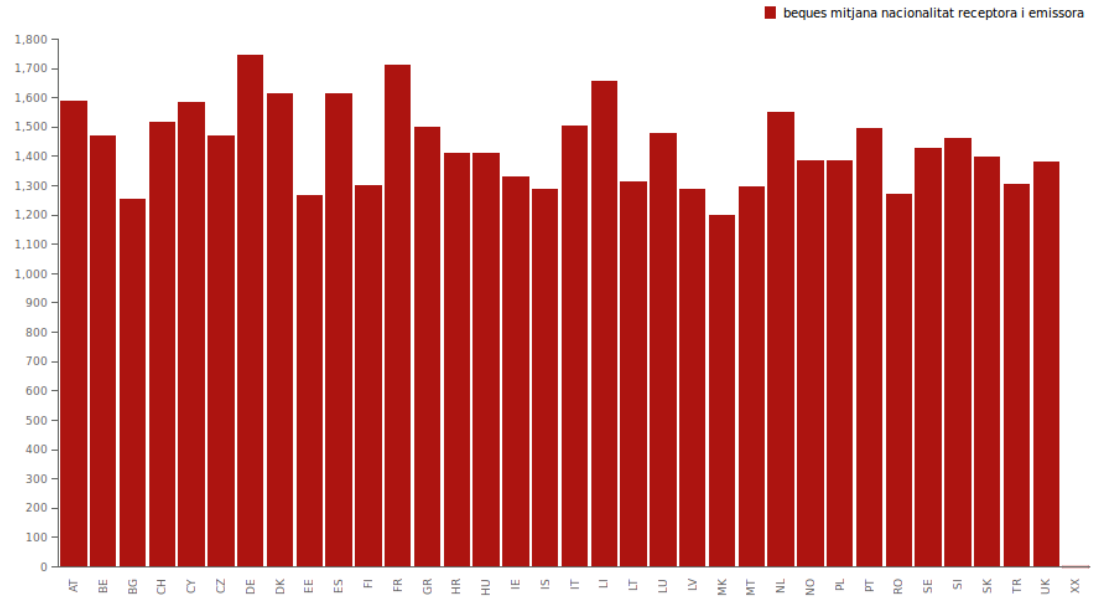
La il·lustració 74 mostra l'informe d'edat mitjana per nacionalitat emissora detall per curs:

	2011	2012
pais_emissor	edat mitjana nacionalitat receptora i emissora	edat mitjana nacionalitat receptora i emissora
AT	23.4	23.4
BE	21.6	21.7
BG	22.5	22.8
CH	23.4	23.5
CY	20.9	21.0
CZ	23.3	23.3
DE	23.4	23.4
DK	23.9	24.0
EE	22.9	23.1
ES	22.7	22.6
FI	23.7	23.7
FR	21.3	21.3
GR	22.2	22.4
HR	23.4	23.2
HU	22.8	22.8
IE	21.6	21.8
IS	25.6	25.3
IT	23.1	23.1
LI	25.0	26.2
LT	21.5	21.5
LU	22.0	22.0
LV	22.4	22.8
MT	21.4	21.2

Il·lustració 74. Edat mitjana per nacionalitat emissora detall

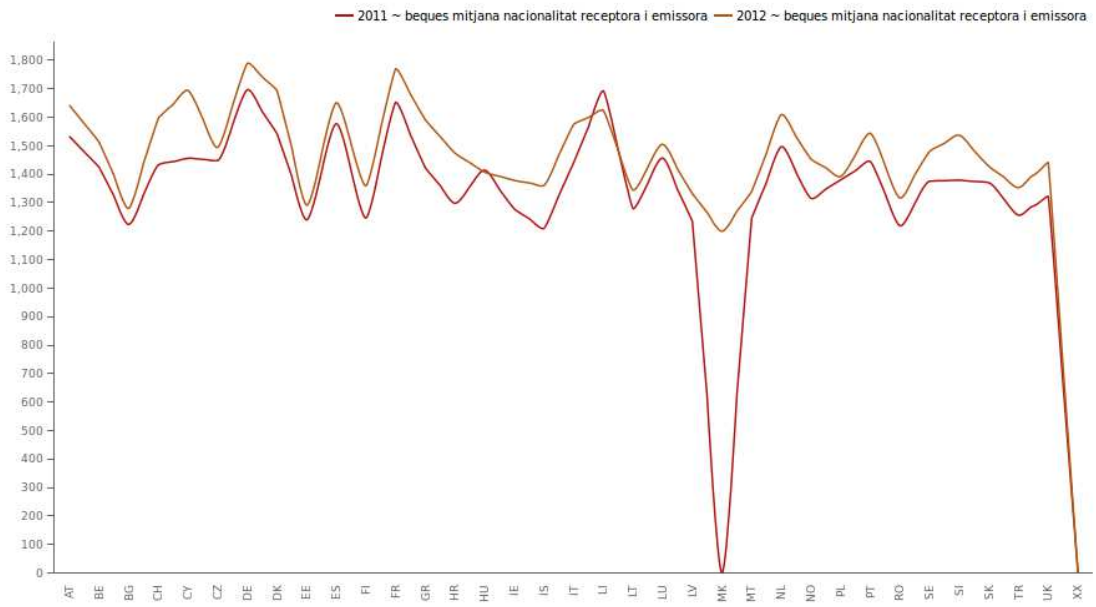
3.4.6 Beques mitjanes per nacionalitat receptora i emissora

La il·lustració 75 mostra l'informe de la mitjana de beques per nacionalitat receptora total:



Il·lustració 75. Mitjana de beques per nacionalitat receptora total

La il·lustració 76 mostra l'informe de la mitjana de beques per nacionalitat receptora detall per curs:



Il·lustració 76. Mitjana de beques per nacionalitat receptora detall

La il·lustració 77 mostra l'informe de la mitjana de beques per nacionalitat emissora total:

pais_emissor	beques mitjana nacionalitat receptora i emissora
AT	1,177.342
BE	1,239.494
BG	2,919.049
CH	1,508.355
CY	2,907.391
CZ	1,322.97
DE	1,258.175
DK	1,179.086
EE	2,566.531
ES	975.668
FI	1,337.45
FR	1,206.464
GR	2,453.3
HR	2,078.881
HU	1,997.541
IE	1,713.926
IS	2,872.469
IT	1,454.056
LI	2,806.025
LT	1,577.422
LU	1,482.602
LV	1,268.324
MT	2,097.108

Il·lustració 77. Mitjana de beques per nacionalitat emissora total

La il·lustració 78 mostra l'informe de la mitjana de beques per nacionalitat emissora detall per curs:

	2011	2012
pais_emissor	beques mitjana nacionalitat receptora i emissora	beques mitjana nacionalitat receptora i emissora
AT	1,098.71	1,254.268
BE	1,222.356	1,255.193
BG	2,791.754	3,039.823
CH	1,737.66	1,290.757
CY	3,152.699	2,727.264
CZ	1,234.318	1,408.039
DE	1,195.572	1,318.036
DK	1,134.426	1,219.692
EE	2,464.037	2,663.602
ES	912.439	1,039.375
FI	1,275.194	1,397.169
FR	1,162.278	1,248.094
GR	2,404.98	2,494.138
HR	1,979.649	2,156.749
HU	1,871.709	2,122.628
IE	1,610.296	1,817.256
IS	2,681.494	3,067.937
IT	1,422.536	1,483.267
LI	2,643.989	3,042.846
LT	1,471.211	1,684.205
LU	1,503.436	1,459.454
LV	1,412.182	1,121.453

Il·lustració 78. Mitjana de beques per nacionalitat emissora detall

3.4.7 Informes addicionals

A la fase de disseny es van proposar dimensions i mesures addicionals per poder construir més informes que els demanats inicialment. A continuació es proposen alguns exemples:

Àrees de coneixement i total de crèdits

Similar a l'informe de mobilitats per àrea de coneixement però s'aprecien diferencies. La Il·lustració 79 mostra l'informe:

descripcio	s_credits
Foreign languages	1,476,182
Business and administration	1,310,291
Business and administration (broad programmes)	1,185,700
Law	985,293
Engineering and engineering trades	607,713
Economics	547,803
Political science and civics	510,364
Humanities	500,932
Medicine	416,826
Architecture and town planning	413,828
Social and behavioural science	319,235
Mother tongue	315,014
Architecture and building	300,678
Management and administration	256,903
Computer science	245,378
Mechanics and metal work	236,661
Biology and biochemistry	233,608

Il·lustració 79. Àrees de coneixement i total de crèdits

Crèdits i mobilitats per gènere

La il·lustració 80 mostra l'informe:

curs	descripcio	s_mobilitats	s_credits
2011	FEMENI	153,468	4,645,176
	MASCULI	99,359	3,091,145
2012	FEMENI	162,962	4,849,179
	MASCULI	104,585	3,198,942
	Grand Total	**520,374**	**15,784,442**

Il·lustració 80. Crèdits i mobilitats per gènere

Evolució temporal de crèdits

A la il·lustració 81 es pot veure l'informe:

curs	s_credits
2011	7,736,321
2012	8,048,121
Grand Total	**15,784,442**

Il·lustració 80. Evolució temporal dels crèdits

Durada mitjana i mobilitats per tipus de mobilitat i gènere

La il·lustració 82 mostra l'informe:

curs	descripcio	descripcio	s_mobilitats	durada mitjana per mobilitat
2011	MIXTA	FEMENI	317	6.248
		MASCULI	121	7.893
	PRACTIQUES	FEMENI	29,365	4.292
		MASCULI	18,718	4.417
	ESTUDI	FEMENI	123,786	6.216
		MASCULI	80,520	6.437
2012	MIXTA	FEMENI	338	6.601
		MASCULI	138	7.25
	PRACTIQUES	FEMENI	34,400	4.276
		MASCULI	21,152	4.383
	ESTUDI	FEMENI	128,224	6.115
		MASCULI	83,295	6.338

Il·lustració 81. Durada mitjana i mobilitats per tipus de mobilitat i gènere

4. Conclusions

El treball final de grau al ser la realització d'un projecte passant per totes les seves fases permet posar en pràctica diferents coneixements i competències que s'han adquirit al llarg de la titulació. Començant per gestió de projectes ja que és un treball extens i és molt important la distribució de les tasques en el temps del que es disposa. Altres assignatures com competència comunicativa per les TIC o anglès també són necessàries per poder documentar i transmetre correctament tot el treball fet per tal que pugui ser entès i avaluat.

Relacionades amb la temàtica del treball les assignatures de l'àrea de bases de dades són la base per entendre que és un magatzem de dades i quins avantatges ens aporta en front les bases de dades operacionals. Per altra banda també s'ha vist que les eines open source són una opció totalment valida alhora de resoldre un projecte amb garanties.

Els objectius a nivell general han estat assolits, s'han posat en pràctica coneixements i competències adquirits durant la titulació. Aquest coneixements i els que s'han adquirit al llarg del projecte de diferents fonts han permès finalitzar el projecte correctament amb la documentació i els productes demanats.

Els objectius a nivell específic també han estat assolits, s'ha entès els avantatges que ens aporten els magatzems de dades en l'ajuda de presa de decisions. També s'ha après a fer un disseny multidimensional i a implementar-lo a partir d'un conjunt de requisits i s'han aportat noves visions als informes. En la fase d'implementació s'han vist els problemes de la integració, transformació i càrrega de dades i s'han resolt per crear un magatzem de dades a partir de múltiples fonts. S'ha generat el cub OLAP per tal de crear tots els informes demanats i s'han proposat informes addicionals.

Pel que fa a la planificació tot i que és difícil planificar un projecte sobre el qual no es tenen els suficients coneixements, en línies generals s'ha seguit el que es va proposar a la primera entrega ja que no hi ha hagut entrebancs greus. S'ha de remarcar però la importància del procés ETL i la inversió d'hores que va suposar. Dels possibles riscos que es van avaluar s'han de tenir en compte el temps que s'inverteix en les correccions que demana el tutor i la inversió de temps que demana aprendre el nou programari.

Per altra banda si es pensa en línies de treball per un futur podrien ser l'anàlisi de magatzem de dades amb eines de mineria de dades en comptes de OLAP. També és interessant la tecnologia Big Data que es diferencia del magatzem de dades ja que no és una arquitectura. Aquestes dues solucions poden arribar a ser compatibles dins d'una mateixa empresa.

Per finalitzar, a nivell personal considero l'experiència molt profitosa i crec que els coneixements adquirits podran servir per projectes futurs que s'imposaran en el meu àmbit laboral com són els de Risk Data Agregation i Risk Reporting.

5. Glossari

TFG: Treball final de grau.

PDF (Portable Document Format): format d'emmagatzematge de tipus compost (imatge vectorial , mapa de bits i text) per a documents digitals independent de plataformes de programari o maquinari.

CSV (comma-separated values): format obert per representar dades en forma de taula, en què les columnes se separen per comes o punt i coma i les files per salts de línia.

XLS: Extensió d'arxiu per defecte del format Excel en versions anteriors o iguals a Excel 2003.

UML (Unified Modeling Language): Llenguatge unificat de modelat, és un llenguatge gràfic per visualitzar, especificar i documentar cadascuna de les parts que comprèn el desenvolupament de programari.

SGBD: Sistema gestor de bases de dades, conjunt de programes que permeten l'emmagatzematge , modificació i extracció de la informació en una base de dades , a més de proporcionar eines per afegir , esborrar, modificar i analitzar les dades.

DW (Data Warehouse): Magatzem de dades.

BI (Business inteligence): Intel·ligència de negoci, consisteix en un conjunt de metodologies, aplicacions i tecnologies que permeten reunir , depurar i transformar dades dels sistemes transaccionals i informació desestructurada en informació estructurada , per a la seva explotació directa (reporting, anàlisi OLTP / OLAP, alertes...) o per a la seva anàlisi i conversió en coneixement , donant així suport a la presa de decisions sobre el negoci.

OLAP (On-line Analytical Processing): Processament analític en línia, solució utilitzada en el camp del BI amb l'objectiu d'agilitzar la consulta de grans quantitats de dades . Per a això utilitza estructures multidimensionals (o cubs OLAP) que contenen dades resumides de grans bases de dades o Sistemes transaccionals (OLTP).

MDX (MultiDimensional eXpressions): Llenguatge de consulta per a bases de dades multidimensionals sobre cubs OLAP .

XML (eXtensible Markup Language): Llenguatge de marques utilitzat per emmagatzemar dades en forma llegible. Dóna suport a bases de dades , sent útil quan diverses aplicacions han de comunicar-se entre si o integrar informació.

ETL (Extract, Transform and Load): Procés que permet moure dades des de múltiples fonts, reformatejar-les, netejar-les i càrregar-les en una altra base de dades, data mart, data warehouse o en un altre sistema operacional per analitzar-les i per donar suport a un procés de negoci.

6. Bibliografia

RIUS, À (COORD.); SERRA, M; ABELLÓ, A; VIDAL, J I CURTO , J. *Data warehouse. Magatzems de dades i models multidimensionals.* Barcelona: Eureca Media (2012)

INMON, W. H. *Building the Data Warehouse* (3a. ed.). EUA: John Wiley & Sons Inc. (2002)

KIMBALL, R; ROSS M. *The Data Warehouse Toolkit* (3a. ed.). Nova York: John Wiley & Sons Inc. (2013)

Pàgines Web consultades al setembre del 2015 :

https://ca.wikipedia.org/wiki/Diccionari_de_dades

http://www.businessintelligence.info/serie-dwh/tablas-de-hecho-fact-tables.html

http://www.dataprix.com/datawarehouse-manager#x1-500003.4.5.1

http://openaccess.uoc.edu/webapps/o2/simple-search?query=data+warehouse

https://wiki.ubuntu.com/Lubuntu

http://lubuntu.net/

https://www.mysql.com/

https://ca.wikipedia.org/wiki/MySQL

http://www.pentaho.com/product/product-overview

http://pentahohispano.blogspot.com.es/2012/01/saiku-la-herramienta-de-analisis-olap.html

http://www.webdetails.pt/ctools/cde/

http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+(Kettle)+Tutorial

http://mondrian.pentaho.com/documentation/workbench.php

https://www.mysql.com/products/workbench/

https://es.wikipedia.org/wiki/Requisito_no_funcional

http://openaccess.uoc.edu/webapps/o2/simple-search?query=data+warehouse

http://infocenter.pentaho.com/help/index.jsp?topic=%2Finstall_graphical%2Ftask_installer_steps.html

http://forums.pentaho.com/showthread.php?161528-Pentaho-Installation-on-Amazon-EC2

http://community.linuxmint.com/tutorial/view/244

Pàgines Web consultades a l'octubre del 2015 :

http://mondrian.pentaho.com/documentation/mdx.php

http://lutarboss4d2.wikispaces.com/Topicos+Avanzados+De+Base+De+Datos

http://4.bp.blogspot.com/-aIaYtvX87TI/VcE7P5g_reI/AAAAAAABGaA/At2sShhBzFk/s1600/Arquitectura_de_Flujo_de_Datos_con_Pentaho_v5_x.png

https://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettlepentaho-data-integration/

https://churriwifi.wordpress.com/2010/05/22/16-3-2-etl-dimension-producto-con-pdi/

http://hdl.handle.net/10609/43024

http://hdl.handle.net/10609/42778

http://hdl.handle.net/10609/43092

Pàgines Web consultades a novembre del 2015 :

https://dev.mysql.com/doc/

http://lutarboss4d2.wikispaces.com/Topicos+Avanzados+De+Base+De+Datos

http://api.ning.com/files/elaRhu1a1S*bHlru-6jfSEy-JXuEpu-INyHuq7ChB7l7ETYRtsP5tAGv8PwjiuRAw125P1LRXBoS6haQRnAV3npbFR31Jsm/PasosparacrearCubosconSchemaWorkbench.pdf

http://www.joyofdata.de/blog/getting-started-with-pentaho-bi-server-5-mondrian-and-saiku/

7. Annexos

Enunciat TFG:

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 un 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.

Els fitxers proporcionats per la Unió Europea són els següents:

- Mobilitat estudiants Erasmus curs 2011-2012 (SM_2011_2012) .
- Mobilitat estudiants Erasmus curs 2012-2013 (SM_2012_2103) .
- Llistat codis i atributs nacionalitats (ISOCountryCodes081507) .
- Llistat codis i atributs institucions (EUC_Consolidated_Table_2007_2013) .
- Llistat codis i atributs àrees de coneixement / especialitzacions (ISCED97_Erasmus_subject_codes).
- Student mobility_2012-2013_Datadictionary .

És recomanable realitzar validacions d'integritat en la informació proporcionada, ja que si bé tota la informació prové de la Unió Europea els sistemes d'origen poden ser diferents.

És important que el Data Warehouse implementat permeti l'actualització dinàmica de totes les dimensions i els fets. Cal poder actualitzar el magatzem de dades amb les dades que es vagin generant en anys consecutius, així com també noves institucions, nacionalitats, etc.

Objectius:

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.

Descripció del treball a realitzar:

L'estudiant rebrà el conjunt de fitxers de Erasmus. A partir d'aquest fitxer i dels requeriments d'usuari esmentats abans, es realitzarà la implementació del magatzem de dades corporatiu. De cara a assolir un correcte desenvolupament del projecte, el construirem per fases o etapes (al final de cada etapa hi haurà un lliurament de PAC en la que s'haurà de lliurar la feina realitzada en aquella fase):

Pla de treball i anàlisi preliminar de requeriments

Al principi del curs es demanarà a l'estudiant un pla de treball on s'indicarà la planificació estimada de les diferents tasques a realitzar per dur a terme el projecte. L'alumne lliurarà, també, un document d'anàlisi preliminar (no detallat) amb l'enumeració i breu descripció dels elements d'anàlisi identificats (dimensions, atributs, indicadors, etc.) que estaran disponibles per als usuaris i el nombre d'informes aproximat que s'implementaran i contingut dels mateixos. També s'analitzaran les fonts de dades operacionals proporcionades que serviran per carregar cadascun dels elements d'anàlisi.

Anàlisi de requeriments i disseny conceptual i tècnic

Es lliurarà un document amb l'anàlisi detallat de requeriments basat en l'anàlisi preliminar realitzat. També es lliurarà un document de disseny amb la descripció del model dimensional que donarà suport a les necessitats dels usuaris, segons l'anàlisi realitzat i el disseny dels procediments d'extracció de dades a alt nivell (processos, pseudocodi, etc.)

Implementació:

Aquesta fase constarà de les tasques següents:

- Construcció del magatzem de dades: base de dades, càrregues, etc.
- Configuració de l'eina d'explotació de dades.
- Construcció dels informes i anàlisi de la informació.

Requeriments de programari:

L'entorn tecnològic estarà format íntegrament per programari lliure:

- Sistema Operatiu: Linux (Lubuntu)
- Base de Dades: Mysql
- Suite BI: Pentaho BA, més el complements Saiku i CDE
- Eines d'ETL: PDI (Kettle)
- Eines Multidimensionals: Mondrian Schema-Workbench
- Eines per el disseny de la base de dades: mysql-workbench

Qualsevol altre programari que es cregui necessari afegir a l'entorn ha de ser validat pel director del Projecte / consultor.