UNIVERSITAT OBERTA DE CATALUNYA

Grau en Enginyeria Informàtica

Construcció i explotació d'un magatzem de dades per a l'anàlisi d'informació sobre allotjaments turístics

Alumne: Víctor Ramón Aixalà Subías

Dirigit per: Carles Llorach Rius

Data: 17 de juny de 2.013

CURS 2012/2013-02

Resum

Aquest document il·lustra el procés dut a terme al hora de realitzar el treball de fi de carrera en els estudis del Grau en Enginyeria Informàtica al àrea de Magatzem de Dades.
Inclou les diferents fases del projecte, des d'una introducció inicial on s'explica que es el *Data Warehouse*, la justificació i objectius del projecte, els requeriments del client, fins el seu anàlisi, disseny i posterior implementació. Aquesta última part ja és de caire molt més tècnic, i depèn de la tecnologia emprada. En aquest cas, la major part del programari utilitzat és de la suite Pentaho.
Caldria ressaltar dos conceptes, que en part són els que causen que el *Data Warehouse* fa anys que sigui vital en la major part d'empreses. D'una banda tenim la integració de dades, donat que en qualsevol empresa existeix una gran dispersió de dades com a resultat dels diferents sistemes utilitzats arreu. D'altra banda la extensió geogràfica de les empreses fa que es trobin ubicades a punts molt llunyans entre si. Per tant, per tenir una visió general de l'empresa és imprescindible aquesta integració de dades que ens brinden els magatzems de dades. A més, aquesta integració de les dades fa possible el segon concepte al qual es volia fer menció, que és l'anàlisi d'aquestes dades per tal d'extreure coneixement útil per la posterior presa de decisions empresarials.

Paraules clau: Magatzems de dades, ETL, MySQL, Pentaho, Turisme, Establiments turístics

Índex de continguts

Taula d'il·lustracions

1. Introducció

Aquest document és la memòria pel Treball de Fi de Grau (TFG) del semestre 2012-2013, al àrea de Magatzem de Dades.

El document inclou una breu descripció del projecte, els requeriments demanats pel client, així com la seva la seva planificació en tasques, un anàlisis de riscos amb el seu pla de contingència, i els dissenys inicials. Després veurem com s'ha realitzat el desenvolupament i si s'ha ajustat o no a les planificacions inicials. Per últim conclourem amb les millores possibles, les línies que es poden desenvolupar i on hi ha més marge de millora.

Començarem indicant per què es creu interessant la realització d'aquest projecte, emmarcant-lo en la situació actual i els objectius que es persegueixen durant el seu desenvolupament.

1.1. Justificació del projecte

Sempre s'ha tingut en consideració de que actualment, a causa de la popularització d'Internet i de forma més recent per l'augment de la seva ubiqüitat per la popularització de l'accés des de dispositius mòbils, tothom es troba envoltat de dades però no es sap treure'n profit d'aquestes en benefici propi. L'interès per aquesta àrea ve de la motivació per aprendre a portar a terme aquest procés d'aprofitament de les dades que doni com a resultat l'obtenció d'informació útil per la presa de decisions.

1.1.1. Què és un Magatzem de dades?

A nivell corporatiu, els DW i les eines OLAP permeten, a partir de l'extracció, transformació i càrrega en el Data Warehouse de les dades emmagatzemades en els sistemes operacionals de la empresa, la explotació en temps real (online) de les esmentades dades permetent oferir suport en la presa de decisions. El magatzem de dades és, per tant, un repositori d'informació orientat a recopilar, resumir i tractar eficientment el gran volum de dades present en les empreses, de manera que faciliti l'anàlisi d'informació des de diverses perspectives o dimensions d'anàlisi, permetent d'aquesta forma la detecció de tendències abans que els competidors, aconseguint així obtenir un avantatge competitiu respecte a ells que permeti realitzar una correcta presa de decisions.

Un magatzem de dades és una base de dades dissenyada més per a l'anàlisi, les consultes de dades i l'expedició d'informes que no pas per als processos de transacció de les dades en sí. Aquesta base de dades permet consolidar dades històriques i derivades de les transaccions de dades. Permet separar la feina d'anàlisi de dades de la feina de recopilació i manteniment de dades.

Un magatzem de dades conté l'estructura d'una base de dades i a més inclou:

> Una eina que permet l'extracció, transformació i càrrega de dades (eina ETL)
>
> Eines per a l'anàlisi en línia de les dades (OLAP)
>
> D'altres aplicacions que permeten als usuaris del magatzem de dades l'emissió de dades analítiques i informes.

Algunes de les seves principals característiques són:

- Orientat als usuaris

- Integrat
- Variable en el temps
- No volàtil
- Conté dades resumides i detallades

Orientat als usuaris

L'objectiu del DW és presentar la informació de la manera més efectiva per la correcta presa de decisions. Per tant, la seva primera característica és que ha de ser útil als usuaris, presentant la informació de forma entenedora.

Integrat

Habitualment, a qualsevol empresa de certa mida existeix una dispersió de les dades. Poden existir diferents aplicacions i bases de dades on hi ha les dades necessàries. Un sistema DW normalitza i homogeneïtza aquest ventall de dades i les emmagatzema en un lloc comú. Es veurà que per això es fan servir les eines ETL (Extracció, Transformació i Càrrega).

Variable en el temps

Es registren els canvis produïts amb el temps, perquè els informes estiguin actualitzats.

No volàtil

La informació continguda al sistema no pateix variacions, ja que només té permisos de lectura. No es poden modificar ni eliminar dades des del sistema DW.

Conté dades resumides i detallades

La informació emmagatzemada té diferents nivells d'agregació de manera que permet navegar pels informes per obtenir més detall d'un punt en concret.
Així, el DW consisteix en el procés de recollir les dades d'un entorn heterogeni i no normalitzat de dades, per normalitzar-lo i extreure informació útil i fiable que permeti al usuari prendre una decisió correcta. Degut al gran volum de dades existent a l'actualitat, els sistemes transaccionals ja no són útils per aquesta tasca. Estem a l'era de l'anomenada "Big Data".
Per tal d'oferir al client tota aquesta informació es fa servir un estructura multidimensional de dades de forma que les consultes d'extracció d'informació siguin òptimes. S'ha intentat reforçar la idea del magatzem de dades des d'un punt de vista funcional respecte el que ofereix al client, i no des de la vessant tecnològica, que malgrat ser força interessant és tan sols un mitjà i no el fi en si mateix.
Amb tot aquestes explicacions queda clara la idoneïtat del projecte, i la seva importància en el món empresarial actual, i en general en qualsevol procés de presa de decisions basada en quantitats ingents de dades. Per això el terme de magatzem de dades també es sol referir com intel·ligència de negoci, que és la capacitat d'extreure coneixement de les dades, on els magatzems de dades són unes de les seves eines principals.

1.2. Objectius

Tal com s'indica al pla docent, 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 uns fitxers proporcionats pel propi client.
A més, el projecte dóna l'oportunitat d'aprendre i perfeccionar la gestió i seguiment de projectes des de les fases inicials, menys la presa de requeriments que en aquest cas ja ve donada. Tot aquest aprenentatge es pot mesurar, ja que el coneixement adquirit es veurà plasmat en la realització d'un producte final, a més d' aquesta memòria i la presentació que l'acompanya.

1.3. Requeriments

A l'enunciat facilitat pel client es veuen els requeriments que ha de complir el producte final. A més, el propi treball de fi de grau (TFG) té uns requeriments addicionals com és la confecció d'aquesta memòria on s'explica el treball realitzat, així com una presentació que resumirà els punts de major interès i transcendència.
L'accés al programari serà a través del portal, on hi haurà penjat els informes requerits pel client. Així doncs, la interfície serà a través d'un web, cosa que permetrà ometre el desenvolupament d'una nova interfície d'usuari i la posterior formació als usuaris en el seu ús.
Cal definir unes proves d'acceptació per tal de validar el sistema, així com els informes, que són la representació final dels requeriments per part del client. D'una banda, aquestes validacions es realitzaran acotant els informes, per tal de poder validar el seus resultats mitjançant el sistema transaccional clàssic.
Finalment, hi ha que preparar el sistema davant futures ampliacions, en concret a l'hora carregar noves dades. També s'han de tenir presents les possibles correccions de les dades en les fases inicials.

1.4. Funcionalitats

Amb el sistema ja en funcionament, aquest ha de ser capaç d'oferir al usuari tota aquesta informació:

- Total d'establiments
- Total de places
- % de places respecte població
- Oferta mitjana de places
- Nombre d'establiments/Nombre d'equipaments
- % de població per equipament
- Indicador d'establiments vs habitants per gènere
- Indicador de places vs persones
- Indicador d'equipaments vs població
- Quantitat de places ofertes / superfície del territori

2. Resultats esperats

S'espera obtenir un producte que compleixi tots els requeriments demanats pel client, amb les seves funcionalitats, complint els temps pactats amb el client. Per tal de controlar el desenvolupament del projecte i que no hagi demores s'estableix una planificació, per tal de poder aplicar els plans de contingència en cas de necessitat. Per fer més controlable aquesta planificació es determinen unes fites intermèdies.

2.1. Planificació inicial vs planificació final

2.1.1. Tasques

S'adjunta una taula amb les tasques i fites principals del projecte, així com la seva planificació i la seva estimació en hores. Tant les tasques com les fites s'han ajustat al calendari pactat amb el client. És a dir, cada fita es correspon amb una entrega al client, menys en el cas de la primera, on es tracta tan sols de recollir la informació necessària pel posterior desenvolupament del projecte.
La segona fita i primera entrega al client inclou aquesta planificació, per pactar amb ell el calendari del projecte i els recursos que s'hi assignaran. Tanmateix s'elabora un pla de contingència per cadascú dels riscos detectats.
Després es realitza un anàlisi i disseny del producte que s'elaborarà. Un cop finalitzat i entregat el disseny, es duu a terme el desenvolupament. Aquest procés inclou la creació de la base de dades, els processos de càrrega i la implementació dels informes que consultarà el client.
Per últim es procedeix a entregar la documentació i presentació al client, així com esperar totes les preguntes que pugui suscitar.

#Tasca	Activitats a realitzar	Inici	Fi	Anterior	Hores
1	**TFG – Data Warehouse**	**27/02/13**	**17/06/13**		-
2	**Fase 1 – Inici TFG**	**27/02/13**	**12/03/13**		-
3	Obtenir recursos del web de la UOC: documentació i materials de l'assignatura[1]	27/02/13	27/02/13		1
4	Lectura i revisió de la documentació i el programari	27/02/13	28/02/13	3	5
5	Instal·lació del programari de suport	03/03/13	03/03/13	4	2
6	**Fase 2 – Pla de Treball TFG – PAC1**	**01/03/13**	**12/03/13**		-
7	Revisió del cas proposat a l'enunciat	01/03/13	01/03/13		1
8	Elaboració del Pla de treball	01/03/13	11/03/13		10
9	Anàlisi preliminar de requeriments de la solució informàtica	02/03/13	11/03/13		5
10	Finalització i Correcció PAC1	03/03/13	11/03/13		2
11	Trobada inicial amb el consultor	10/03/13	10/03/13		0
12	Fita 1 – Lliurament PAC1	12/03/13	12/03/13	10	-
13	**Fase 3 – Anàlisi i Disseny de la solució informàtica – PAC2**	**13/03/13**	**16/04/13**		-
14	Estudi conceptes nous relacionats amb DW: eines ETL, OLAP	13/03/13	22/03/13		8
15	Anàlisi de requeriments de la solució informàtica	19/03/13	25/03/13		30
16	Disseny de la base de dades	26/03/13	03/04/13	14	4
17	Disseny de les eines ETL i OLAP	04/04/13	10/04/13	15	0
18	Disseny informes	11/04/13	14/04/13	16	0
19	Instal·lació de Màquina Virtual del projecte	13/04/13	14/04/13		8
20	Revisió PAC2	15/04/13	15/04/13		1
21	Fita 2 – Lliurament PAC2	16/04/13	16/04/13	19	-
22	**Fase 4 – Implementació del magatzem de dades – PAC3**	**17/04/13**	**29/05/13**		-
23	Creació física de la base de dades	17/04/13	02/05/13		1
24	Implementació dels processos ETL	03/05/13	10/05/13	22	20
25	Implementació dels processos OLAP	11/05/13	18/05/13	23	0
26	Implementació dels informes	19/05/13	26/05/13	24	40

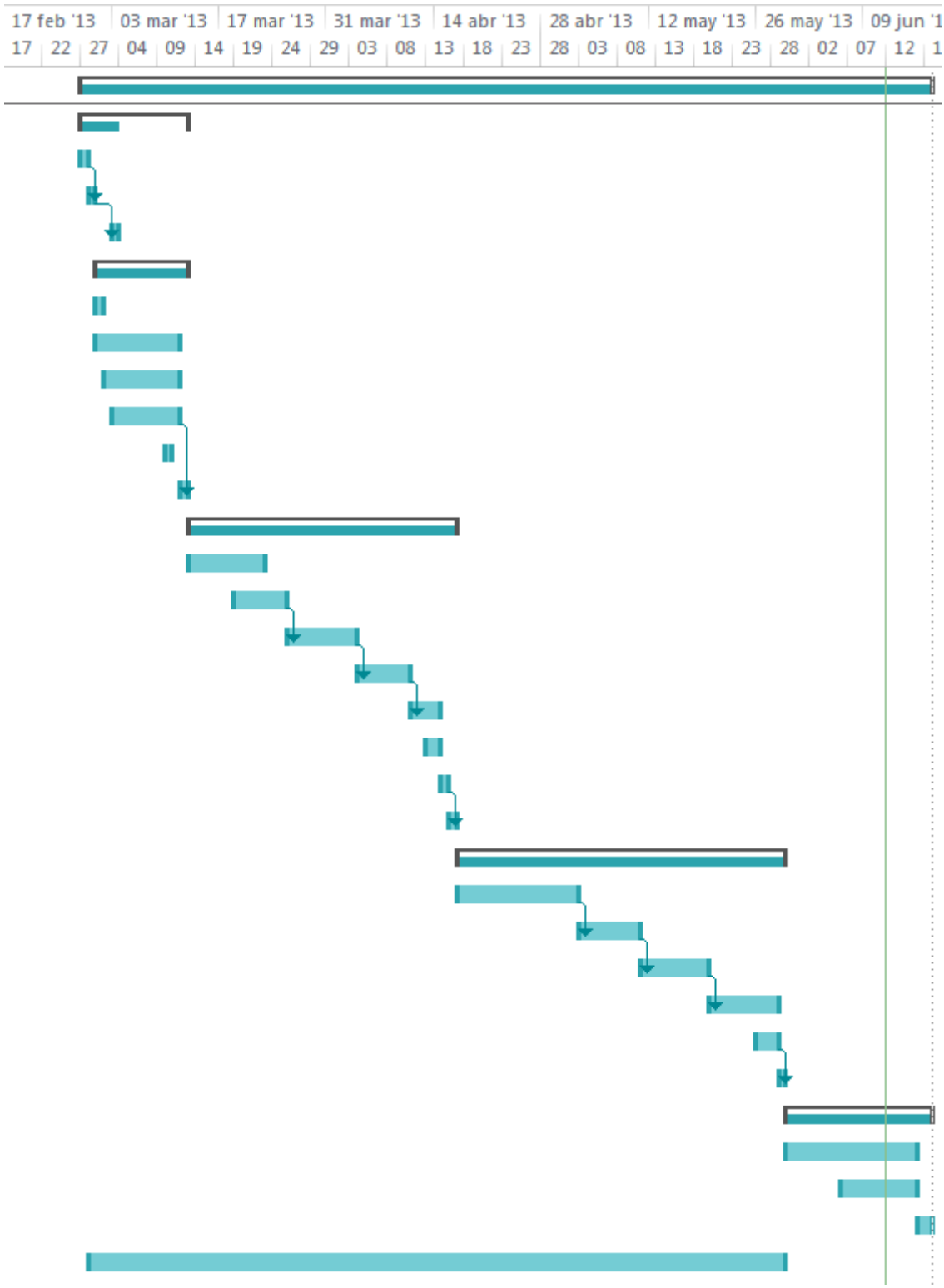
27	Revisió PAC3	26/05/13	29/05/13		1
28	Fita 3 – Lliurament PAC3	29/05/13	29/05/13	26	-
29	**Fase 5 – Lliurament final. Memòria i presentació virtual**	**30/05/13**	**17/06/13**		-
30	Elaboració de la memòria final	30/05/13	15/06/13		10
31	Elaboració de la presentació virtual	06/06/13	15/06/13		
32	Fita 4 – Revisió i Lliurament Final	16/06/13	17/06/13		-
33	Recopilació de dades i documentació per a la memòria del projecte	28/02/13	29/05/13		0

Il·lustració 1: Taula de tasques

2.1.2. Diagrama de Gantt

	Nombre de tarea	Comienzo	Fin
1	**TFG - Data Warehouse**	**mié 27/02/13**	**lun 17/06/13**
2	**Fase 1 - Inici TFG**	**mié 27/02/13**	**mar 12/03/13**
3	Obtenir recursos del web de la UOC: documentació i materials de l'assignatura	mié 27/02/13	mié 27/02/13
4	Lectura i revisió de la documentació i el programari	jue 28/02/13	jue 28/02/13
5	Instal·lació del programari de suport	dom 03/03/13	dom 03/03/13
6	**Fase 2 - Pla de Treball TFG - PAC1**	**vie 01/03/13**	**mar 12/03/13**
7	Revisió del cas proposat a l'enunciat	vie 01/03/13	vie 01/03/13
8	Elaboració del Pla de Treball	vie 01/03/13	lun 11/03/13
9	Anàlisi preliminar de requeriments de la solució informàtica	sáb 02/03/13	lun 11/03/13
10	Finalització i Correcció PAC1	dom 03/03/13	lun 11/03/13
11	Trobada inicial amb el consultor	dom 10/03/13	dom 10/03/13
12	Fita 1 - Lliurament PAC1	mar 12/03/13	mar 12/03/13
13	**Fase 3 - Anàlisi i Disseny de la solució informàtica - PAC2**	**mié 13/03/13**	**mar 16/04/13**
14	Estudi conceptes nous relacionats amb DW: eines ETL, OLAP	mié 13/03/13	vie 22/03/13
15	Anàlisi de requeriments de la solució informàtica	mar 19/03/13	lun 25/03/13
16	Disseny de la base de dades	mar 26/03/13	mié 03/04/13
17	Disseny de les eines ETL i OLAP	jue 04/04/13	mié 10/04/13
18	Disseny informes	jue 11/04/13	dom 14/04/13
19	Instal·lació de Màquina Virtual del projecte	sáb 13/04/13	dom 14/04/13
20	Revisió PAC2	lun 15/04/13	lun 15/04/13
21	Fita 2 - Lliurament PAC2	mar 16/04/13	mar 16/04/13
22	**Fase 4 - Implementació del magatzem de dades - PAC3**	**mié 17/04/13**	**mié 29/05/13**
23	Creació física de la base de dades	mié 17/04/13	jue 02/05/13
24	Implementació dels processos ETL	vie 03/05/13	vie 10/05/13
25	Implementació dels processos OLAP	sáb 11/05/13	dom 19/05/13
26	Implementació dels informes	lun 20/05/13	mar 28/05/13
27	Revisió PAC3	dom 26/05/13	mar 28/05/13
28	Fita 3 - Lliurament PAC3	mié 29/05/13	mié 29/05/13
29	**Fase 5 - Lliurament final. Memòria i presentació virtual**	**jue 30/05/13**	**lun 17/06/13**
30	Elaboració de la memòria final	jue 30/05/13	sáb 15/06/13
31	Elaboració de la presentació virtual	jue 06/06/13	sáb 15/06/13
32	Fita 4 - Revisió i Lliurament Final	dom 16/06/13	lun 17/06/13
33	Recopilació de dades i documentació per a la memòria del projecte	jue 28/02/13	mié 29/05/13

Il·lustració 2: Planificació inicial



Il·lustració 3: Diagrama de Gantt

2.1.3. Planificació Final

La planificació s'ha anat complint fins la fita d'entrega de la PAC2. En la realització de la PAC3, degut a que inicialment la càrrega de dades s'havia fet manualment i no s'havien confeccionat processos ETL, a més de la dificultat inherent per la implementació dels informes amb un programari que mai s'havia utilitzat, s'han trobat més problemes dels planificats per complir els temps que s'havien planificat inicialment.
De forma més concreta, a l'apartat corresponent al procés ETL s'ha donat un problema, l'aprenentatge d'una nova eina, ja que l'eina utilitzada no era del tot intuïtiva respecte el que hom desitjaria. Aquest aspecte s'ha solucionat mitjançant el wiki de la pròpia suite Pentaho.
Posteriorment, a la fase d'implementació dels informes s'ha creat un camp compost pel sumatori de poblacions en aquells informes que es combinés amb establiments o equipaments, el qual ha ralentitzat de forma notable la generació dels informes on s'involucressin el nombre d'equipaments davant la gran cardinalitat d'aquesta dimensió. Tot i no suposar un repte al desenvolupament dels informes, això ha suposat un contratemps a l'hora d'avaluar aquests informes cada vegada que es llençava una prova.
Això deixa palès que caldria haver previst en la planificació inicial més temps per familiaritzar-se amb aquestes eines, i haver creat una tasca pel aprenentatge i parametrització de totes elles.

2.2. Anàlisis de riscos

En aquest apartat es veuran els possibles riscos detectats i el seu pla de contingència.

2.2.1. Equips Informàtics

El principal problema i risc al que es fa front durant el desenvolupament de qualsevol tasca informàtica es que un equip o perifèric com ara el disc dur que es fa servir es faci malbé.
Les causes per les que això pot passar són molt diverses, i per mencionar unes quantes es poden destacar una pujada de tensió, apagar l'ordinador sense fer servir els passos correctes, un virus, etcètera...
Per tant, s'haurà de tenir una còpia de seguretat actualitzada de forma freqüent per poder estar més tranquils. Aquesta còpia es farà en un dispositiu extern tal com un disc dur extern o una memòria USB. A més, es guardaran copies dels documents a la xarxa, com ara al correu electrònic o a un servei d'emmagatzematge a Internet com *Dropbox.*

2.2.2. Altres assignatures

El fet d'estar matriculat de varies assignatures pot fer que un endarreriment en una assignatura afecti la planificació del projecte. És acceptable un petit endarreriment en la planificació, però en casos més greus s'ha de donar prioritat al projecte.

2.2.3. Imprevistos laborals

Un altre punt a tenir en compte és que un projecte laboral provoqui que es tinguin que abocar-hi més esforços, en cas que l'autor del projecte trobi feina. Caldrà llavors realitzar una nova planificació del projecte, i advertir al consultor.

2.2.4. Greus endarreriments

Es pot donar el cas que durant el desenvolupament del treball de fi de carrera es detecti un greu

endarreriment, que pot ser provocat tant per una mala planificació inicial com per un problema sorgit durant el treball amb el que no es comptava. En aquest cas, caldrà advertir al consultor del problema i planificar-ho de nou el més aviat possible perquè reflecteixi la situació actual. També pot ocasionar dedicar més recursos dels inicialment previstos al projecte.

2.2.5. Aprenentatge Eines

Qualsevol eina nova comporta al principi una corba d'aprenentatge més plana, amb pocs progressos pel temps invertit. En aquest projecte existeixen moltes eines, on la major part d'elles es treballa per primer cop amb elles, per la qual cosa aquest risc ha de ser tingut en compte. Per combatre'l tan sols hi ha dues maneres, dedicar-hi més temps del previst, o fer ús de la bibliografia o dels articles a Internet, per tal d'accelerar l'aprenentatge.

3. Anàlisi i disseny

3.1. Requeriments funcionals

L'empresa Observatori Nacional d'Ocupació (ONdO) vol crear un magatzem de dades sobre el nombre d'establiments turístics per tal de poder aprofundir en l'evolució d'aquest tipus d'establiments que ofereixen gairebé sis-centes mil places a Catalunya, i analitzar les possibles correlacions entre allotjaments i equipaments públics.

Concretament es demanen una sèrie d'informes:

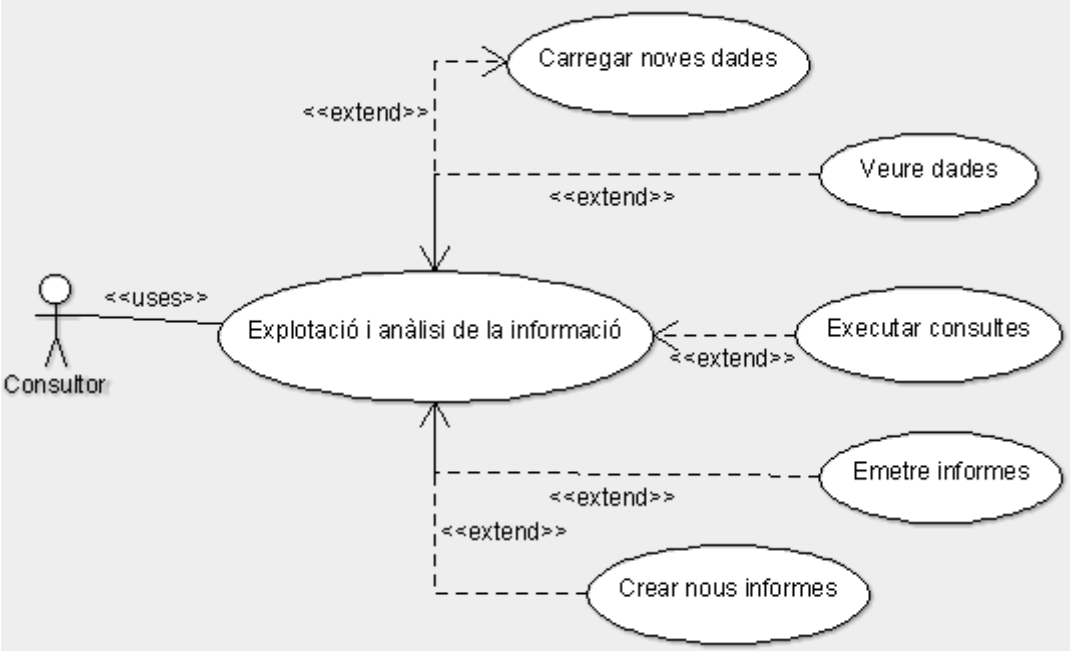
- **Total d'establiments:** un llistat amb el nombre d'establiments d'un determinat tipus per una demarcació determinada i un any donat.
- **Total de places:** un llistat amb el nombre de places d'un determinat tipus per una demarcació determinada i un any donat.
- **% de places respecte població:** llistat comparatiu entre el nombre de places d'un determinat tipus i el cens de població per una demarcació determinada i un any donat.
- **Oferta mitjana de places:** nombre mitjà de places d'un tipus determinat durant un període comprés entre dos anys per una demarcació determinada.
- **Nombre d'establiments/Nombre d'equipaments:** llistat comparatiu entre el nombre d'establiments i equipaments per una demarcació determinada i un any donat.
- **% de població per equipament:** relació percentual entre el cens d'una demarcació per un any donat i el nombre d'equipaments per aquesta demarcació.
- **Indicador d'establiments vs habitants per gènere:** mesura en forma de ràtio que dóna una relació entre el nombre d'establiments i el cens per gènere per una demarcació determinada i un any donat.
- **Indicador de places vs persones:** mesura en forma de ràtio que dóna una relació entre el nombre de places i el cens per una demarcació determinada i un any donat.
- **Indicador d'equipaments vs població:** mesura en forma de ràtio que dóna una relació entre el nombre d'equipaments i el cens per una demarcació determinada i un any donat.
- **Quantitat de places ofertes / superfície del territori:** mesura en forma de ràtio que dóna una relació entre el nombre de places i la superfície d'un territori per una demarcació determinada i un any donat.

Tanmateix ja s'indica amb quin agrupament i en quines dimensions es voldrà aquest informació. La temporalitat serà anual, i les dades es podran consultar de forma agregada, per comarca / província, tipus d'establiment i categoria. De totes formes, als propis informes ja s'indiquen les dimensions per les quals s'està interessat veure la informació.

3.1.1. Casos d'ús

Inicialment, es va adoptar la filosofia de que hi havia dos actors, el consultor que fa les seves tasques d'anàlisi de les dades i utilitza els informes dissenyats per l'aplicació, i l'administrador encarregat de gestionar l'aplicació, generar els nous informes que li demanin els analistes i fer la càrrega de noves dades a mesura que aquestes s'hagin d'actualitzar.
Posteriorment, s'ha apreciat que l'automatització dels processos exclou la necessitat de tindre un perfil d'administrador. D'una banda, s'ha dut a terme una automatització dels processos ETL a mesura que s'incorporin noves dades. Per tant, aquest cas d'ús es pot ometre del perfil d'administrador ja que serà el mateix consultor qui dugui a terme aquesta tasca. D'altra banda, algunes de les eines del Pentaho BI Server

permeten que el consultor dissenyi els seus propis informes amb un cert grau de personalització. Conseqüentment, aquesta tasca tampoc requereix de la presència d'un perfil d'administrador. Aquest nou plantejament fa que en cas que s'hagin de dur a terme canvis al sistema es realitzi un nou projecte DataWarehouse. En conclusió, es pot afirmar que sota aquest nou plantejament la presència d'un perfil d'administrador no es considera necessària. El diagrama definitiu de casos d'ús pel consultor seria aquest.



Il·lustració 4: Diagrama de casos d'ús

3.2. Requeriments No Funcionals

En aquest cas es volia una aplicació àgil, on el usuari pogués consultar els informes indicats, amb unes prestacions de rendiment correctes, per la qual cosa s'havia de fer un estudi del maquinari adient, segons les necessitats que es determinessin.
Aquesta eina ha de permetre l'accés concurrent per diferents usuaris. Hi ha un únic tipus d'usuari, el consultor, que tindrà permisos per a l'explotació i anàlisi de les dades que permeti extreure'n informació útil d'elles. Aquesta tasca inclou les tasques per a carregar noves dades mitjançant processos automatitzats si és necessari, consultar les dades ja sigui de forma natural o mitjançant informes personalitzats i demanats pel propi client, com generar nous informes amb un cert grau de personalització si així ho requereix.
Actualment no està previst restringir informes segons el usuari, però a petició del client es podria estudiar. Es tractaria que un usuari X només tingui accés a una sèrie d'informes, adequats a la seva tasca, i no tingués accés a la resta d'informes perquè no afecten al seu camp professional.
Al tenir tota la informació en una base de dades estàndard com MySQL, no serà costós executar una portabilitat, de fet el propi Spoon de Pentaho, que es fa servir pel procés ETL, es pot utilitzar per realitzar una migració.
A més, durant l'elaboració del projecte es generarà tota una documentació, que en aquest cas seran les diverses entregues i aquesta memòria final.

3.3. Model Conceptual

Segons els requeriments esmentats es poden obtenir les mesures i dimensions que a priori ja es poden establir. Després es veurà la taula de fets, així com el seu nivell de detall i granularitat.

3.3.1. Mesures

En aquest cas s'han pogut contemplar les següents mesures:

- Nombre d'establiments: nombre d'establiments d'un determinat tipus per una comarca determinada a un any donat.
- Nombre de places: nombre de places d'un determinat tipus per una comarca determinada a un any donat.
- Nombre d'equipaments: nombre d'equipaments d'un determinat tipus per una comarca determinada a un any donat.
- Població: cens d'habitants per una comarca determinada a un any donat. Per indicació expressa del client es calcularà com la mitjana entre el cens de població de l'any en curs i el següent. Per tant, el valor per la població al 2.011 serà el resultat de fer la mitjana entre el cens pels anys 2.011 i 2.012. En cas que encara no es disposessin de les dades pel any següent perquè són de l'any en curs (actualment seria el cens pel 2.013 que demanaria els censos dels anys 2.013 i 2.014), temporalment el cens per l'any en curs seria el cens a 1 de gener del mateix any.

La resta de mesures, com ràtios i percentatges, seran calculades posteriorment en temps d'execució. Totes aquestes altres mesures no es poden incorporar directament ja que són calculades i no agregables.

3.3.2. Dimensions

Per aquest projecte es poden considerar quatre dimensions:

- Temps: Període de validesa, amb periodicitat anual, sobre les dades que s'emmagatzemen a la taula del fet.
- Demarcació: Dades sobre una comarca, com la superfície que ocupa i una proporció de la seva població per gènere a l'any 2.012 que s'utilitza per extrapolar dades de població per gènere per anys anteriors a l'informe que requereix aquestes dades.
- Establiment: classificació per grup i categoria dels diversos tipus d'establiments turístics que es poden emmagatzemar al sistema.
- Equipament: classificació per grup i categoria dels diversos tipus d'equipaments públics que es poden emmagatzemar al sistema.

3.3.3. Taula de Fets

La taula de fets, a nivell de granularitat, es correspon amb la informació del nombre d'establiments i places, de forma que per aquesta dimensió no es fa cap mena d'agrupació prèvia. D'altra banda, respecte a la informació per demarcació s'ha tingut que agrupar per comarques ja que els fitxers proporcionats pel client tenien una granularitat a nivell de població, i la informació per província es calcula en temps d'execució com agregació de les seves comarques. Finalment, la informació pel nombre equipaments té caràcter permanent al llarg de tots els anys considerats ja que no es proporciona informació amb periodicitat anual

per aquesta dimensió. Per saber el volum de dades al qual es fa front es poden agrupar les dades per any.

Any	Nombre de registres
2006	19.962
2007	22.180
2008	22.180
2009	22.180
2010	22.180
2011	22.180
2012	22.180
TOTAL	**153.042**

Es tracta d'un nombre força considerable. Conseqüentment, el disseny de la taula de fets és aquest:

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	Default
id	INT(10)	☑	☑	☑	☐	☑	☐	☑	
id_poblacio	INT(11)	☐	☑	☐	☐	☐	☐	☐	
temps	INT(11)	☐	☑	☐	☐	☐	☐	☐	
id_cat_establ	INT(11)	☐	☑	☐	☐	☐	☐	☐	
id_cat_equip	INT(11)	☐	☑	☐	☐	☐	☐	☐	
num_establiments	DOUBLE	☐	☐	☐	☐	☐	☐	☐	NULL
num_places	DOUBLE	☐	☐	☐	☐	☐	☐	☐	NULL
num_equipaments	INT(11)	☐	☑	☐	☐	☐	☐	☐	
poblacio	DOUBLE	☐	☑	☐	☐	☐	☐	☐	
		☐	☐	☐	☐	☐	☐	☐	

Il·lustració 5: Taula de Fets

3.3.4. Origen de les dades

Inicialment es va proporcionar un fitxer comprimit amb els següents arxius:

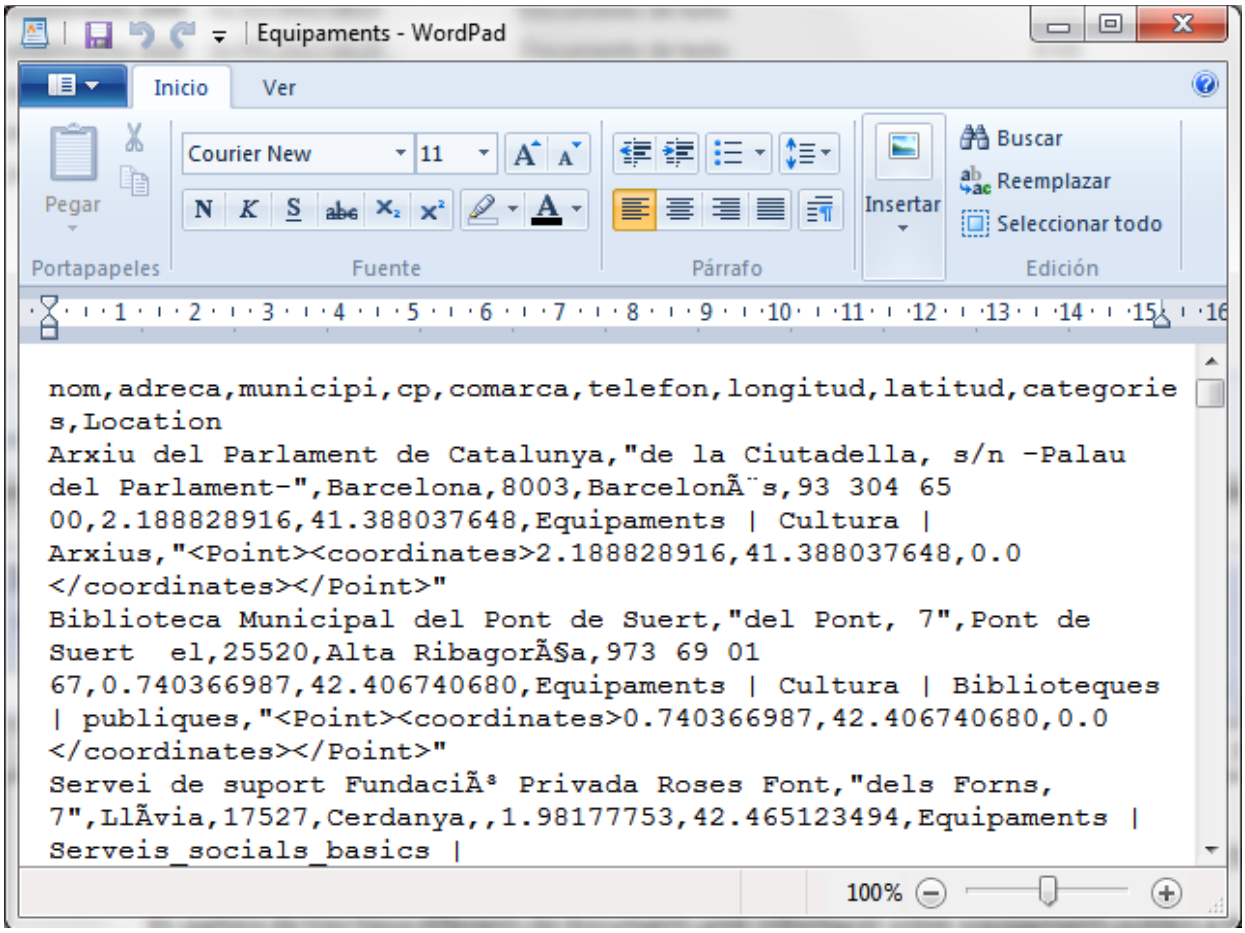
Nombre	Fecha de creación	Tipo	Tamaño
Equipaments	01/03/2013 18:25	Full de càlcul de l'OpenDocument	8.765 KB
establiments 2006	01/03/2013 18:25	Documento de texto	9 KB
establiments 2007	01/03/2013 18:25	Documento de texto	9 KB
establiments 2008	01/03/2013 18:25	Documento de texto	9 KB
establiments 2009	01/03/2013 18:25	Documento de texto	9 KB
establiments 2010	01/03/2013 18:25	Documento de texto	9 KB
establiments 2011	01/03/2013 18:25	Documento de texto	10 KB
establiments 2012	01/03/2013 18:25	Documento de texto	14 KB
poblacio	01/03/2013 18:25	Full de càlcul de l'OpenDocument	37 KB

Il·lustració 6: Fitxers proporcionats pel client

Es van catalogar aquests arxius en tres tipus:

Arxiu d'equipaments:

Un arxiu .csv amb un llistat dels diversos equipaments públics existents a Catalunya a data 31/12/2012.



Il·lustració 7: Fitxer d'equipaments

Els camps dels quals es composa aquest fitxer són:

Nom: en aquest camp hi ha el nom de l'equipament públic.

Adreça: direcció on es troba situat l'equipament públic.

Municipi: municipi on es troba l'equipament públic, amb els articles inicials que pot haver-hi a una població al final del camp.

Codi postal: codi postal on es troba situat l'equipament públic.

Comarca: comarca on es troba situat l'equipament públic, en aquest cas normalitzat. O sigui, la primera lletra de cada paraula en majúscules i la resta en minúscules.

Telèfon: telèfon de contacte de l'equipament públic.

Longitud: coordinada de longitud on es troba l'equipament públic.

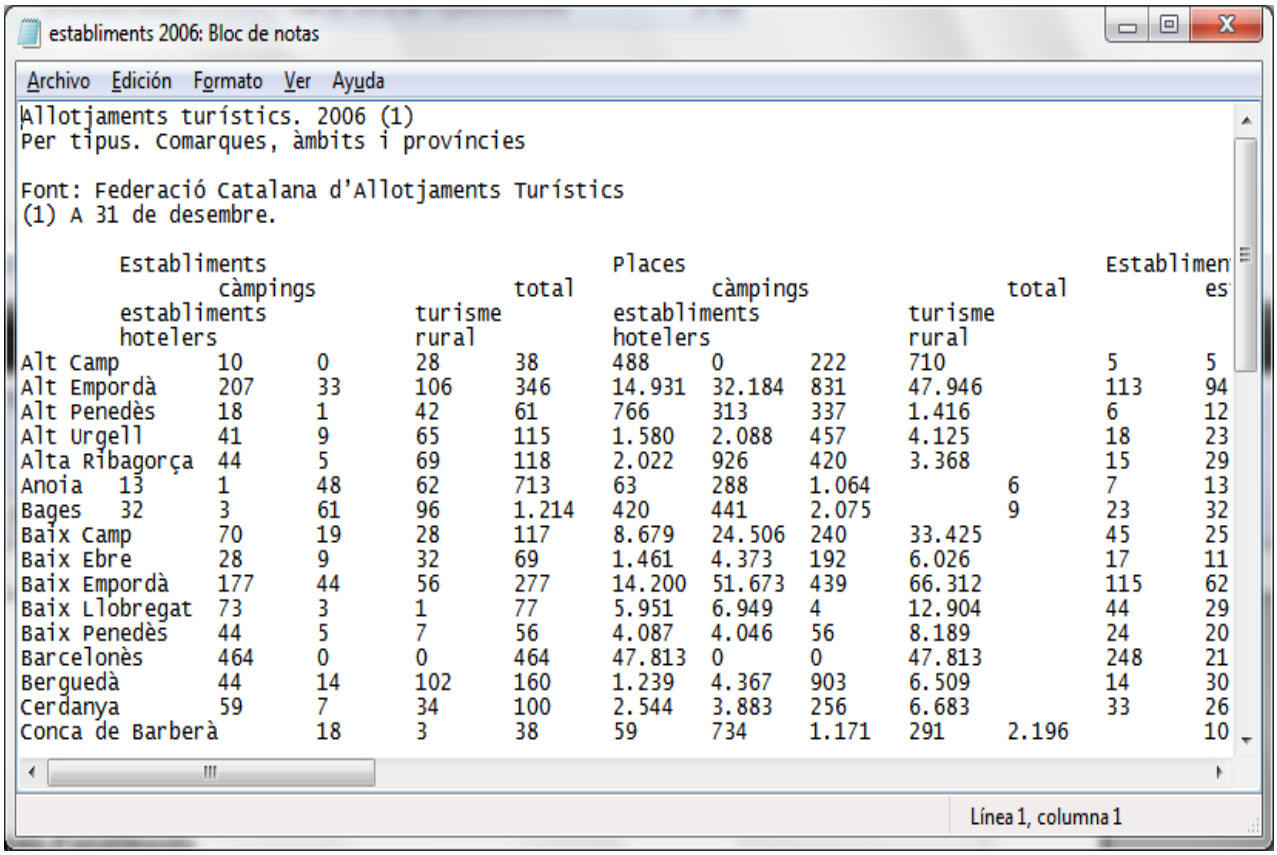
Latitud: coordinada de latitud on es troba l'equipament públic.

Categoria: categoria, amb un total de tres nivells, a la que pertany aquest equipament públic. Tots els nodes parteixen del node arrel "Equipaments".

Localització: informació redundant respecte als camps de longitud i latitud, que es suposa que és d'utilitat a l'hora de situar un punt a un programari com Google Maps.

Arxius d'establiments:

Set arxius .txt repartits entre els anys 2006 i 2012 amb un llistat dels diferents establiments que es poden trobar a Catalunya.



Allotjaments turístics. 2006 (1)
Per tipus. Comarques, àmbits i províncies

Font: Federació Catalana d'Allotjaments Turístics
(1) A 31 de desembre.

	Establiments				Places					
	càmpings			total	càmpings			total		
	establiments hotelers		turisme rural		establiments hotelers		turisme rural			
Alt Camp	10	0	28	38	488	0	222	710	5	5
Alt Empordà	207	33	106	346	14.931	32.184	831	47.946	113	94
Alt Penedès	18	1	42	61	766	313	337	1.416	6	12
Alt Urgell	41	9	65	115	1.580	2.088	457	4.125	18	23
Alta Ribagorça	44	5	69	118	2.022	926	420	3.368	15	29
Anoia	13	1	48	62	713	63	288	1.064	6	7
Bages	32	3	61	96	1.214	420	441	2.075	9	23
Baix Camp	70	19	28	117	8.679	24.506	240	33.425	45	25
Baix Ebre	28	9	32	69	1.461	4.373	192	6.026	17	11
Baix Empordà	177	44	56	277	14.200	51.673	439	66.312	115	62
Baix Llobregat	73	3	1	77	5.951	6.949	4	12.904	44	29
Baix Penedès	44	5	7	56	4.087	4.046	56	8.189	24	20
Barcelonès	464	0	0	464	47.813	0	0	47.813	248	21
Berguedà	44	14	102	160	1.239	4.367	903	6.509	14	30
Cerdanya	59	7	34	100	2.544	3.883	256	6.683	33	26
Conca de Barberà		18	3	38	59	734	1.171	291	2.196	10

Il·lustració 8: Fitxer d'establiments

Dins d'aquests arxius es poden observar que hi ha el nombre d'establiments i places detallat en diferents tipus i subtipus, a nivell de comarca i província, a més dels totals agregats per cada tipus. S'han dut a terme diversos canvis de tipologia durant aquests anys. Els tipus i subtipus en que s'han dividit els arxius són aquests:

Establiments
- Establiments hotelers
  - Establiments estrelles d'or
  - Establiments estrelles d'argent

A l'any 2012 es van canviar aquestes dues tipologies:

- Hotels
- Hostals o pensions

- Càmpings
  - Establiments de 1ª
  - Establiments de 2ª
  - Establiments de 3ª
  - Establiments privats en l'any 2006
  - Establiments de luxe en l'any 2007 i posteriors

- Turisme rural

  - A l'any 2006 els establiments es dividien en:
    - Allotjament independent
    - Masia
    - Casa de poble

  - A l'any 2007 i posteriors van passar a dividir-se en:
    - Casa de poble compartida
    - Casa de poble independent
    - Masia
    - Masoveria

Respecte els diversos canvis de tipologia al llarg dels anys analitzats s'han pres les següents decisions:

- No hi ha equivalència entre els diversos tipus d'hotels abans de 2012 i a partir de 2012, ja que la harmonia a les dades no és evident.
- Hi ha equivalència entre els càmpings privats i els càmpings de luxe, ja que la cardinalitat és la mateixa i la harmonia a les dades és bastant evident.
- No hi ha equivalència entre els diversos tipus d'establiment de turisme rural abans i després de 2007, per qüestions de canvi en la cardinalitat dels dos conjunts.

Arxiu de poblacions:

	A	B	C	D	E	
1	Municipi	Codi INE	Població 2012	Població 2011	Població 2010	P
2	Abrera	8001	11611	11469	11521	
3	Agramunt	25003	5653	5618	5608	
4	Aiguafreda	8014	2481	2505	2464	
5	Aitona	25038	2405	2393	2398	
6	Albatàrrec	25007	2040	1979	1872	
7	Albesa	25008	1656	1679	1613	
8	Albinyana	43002	2347	2314	2275	
9	Alcanar	43004	10601	10545	10570	
10	Alcarràs	25011	8350	8029	7776	
11	Alcoletge	25012	3015	2816	2677	
12	Alcover	43005	5140	5143	5100	
13	Aldea (L')	43904	4513	4376	4063	

Il·lustració 9: Fitxer de poblacions

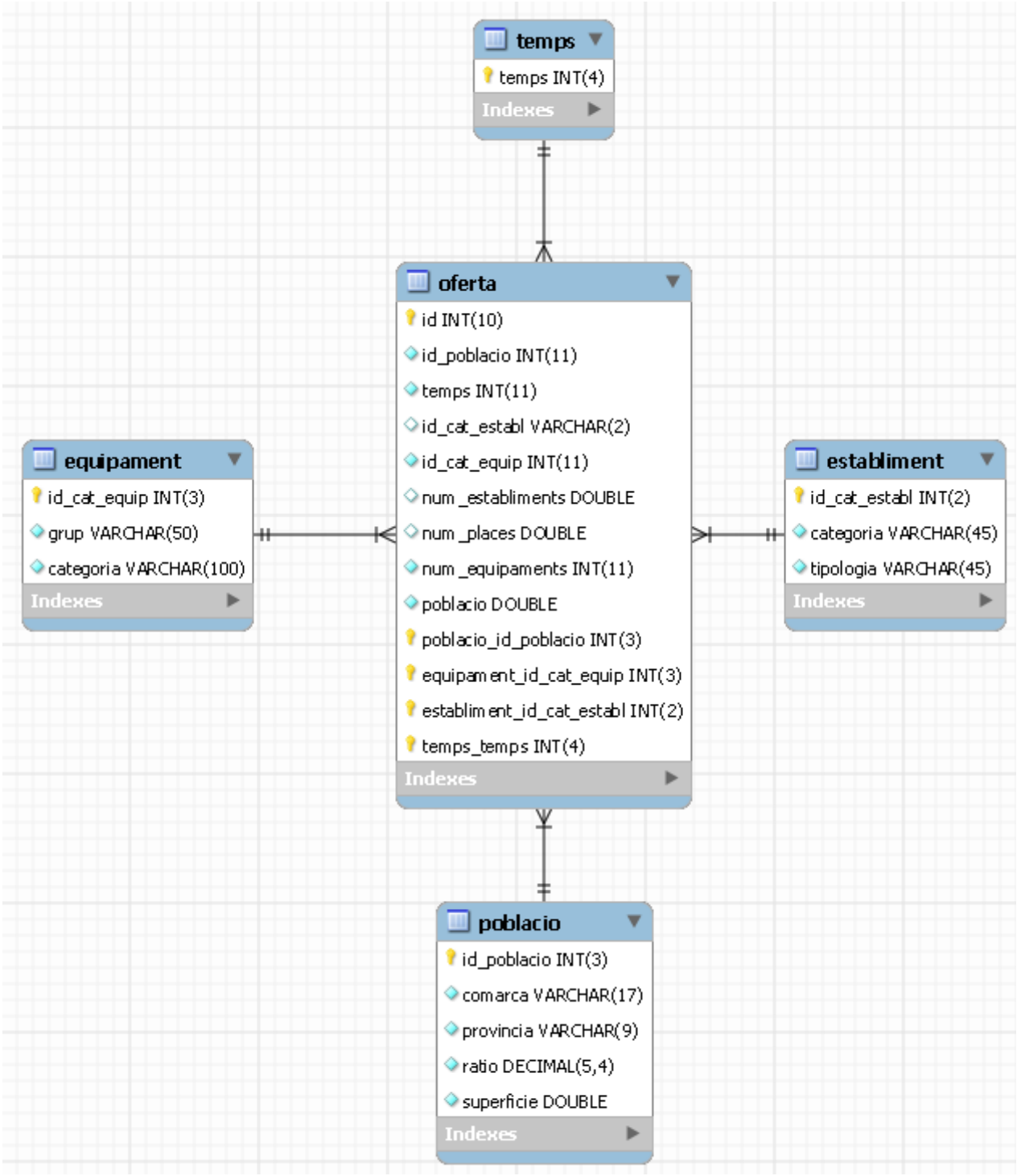
Un arxiu .csv amb registre anual del nombre d'habitants des de l'any 2006, detallat per sexe per l'any 2012. Donat que el cens de població es calcularà com la mitjana entre un any i el següent, el cens per l'últim any, davant la impossibilitat de tindre dades per l'any següent, serà el cens a 1 de gener de l'any actual. Concretament, el cens per l'any 2011 seria la mitjana del cens als anys 2011 i 2012. I el cens per l'any 2012 seria el cens a 1 de gener de 2012. A més, respecte al fet que hi ha un requisit que demana un indicador del nombre d'establiments respecte els habitants per gènere, i que no hi ha dades detallades per gènere per abans del 2012, es seguirà la recomanació del consultor d'aplicar el percentatge d'homes/dones respecte l'any 2012 sobre el total dels anys anteriors.

3.3.5. Base de dades temporal (Staging Area)

Aquestes dades d'orígens diferents es carregaran a una base dades del servidor MySQL per poder manipular-les més fàcilment. A aquesta base de dades temporal se l'anomena *Staging Area*, i serveix per fer les transformacions necessàries abans de fer la càrrega al magatzem de dades.

L'estructura és bastant diferent al contingut dels fitxers d'origen. En primer lloc, la taula de temps s'omple manualment davant la baixa quantitat de registres que té (un per cada any). La taula d'establiments també s'omple manualment, ja que no conté informació numèrica i només hi ha un catàleg dels diversos tipus d'establiments que es poden trobar a la taula dels fets. El mateix es podria fer amb la taula d'equipaments tractant-se d'un catàleg de les diverses categories d'equipaments que es poden trobar reflexades al sistema, però en aquest cas la cardinalitat del conjunt demana algun tipus d'automatització en la càrrega de dades. Per aquesta càrrega només s'agafarà el grup al qual pertany un equipament (cultura, educació, mobilitat i transports...) i les categories incloses dintre d'aquest grup del fitxer d'equipaments, ja que la resta de dades del fitxer d'equipaments proporcionat pel client no són d'interès pel sistema plantejat. Finalment, la informació del cens i superfície per les diverses poblacions catalanes s'agrupa per comarques, i s'afegeix la província a cada registre seguint el plantejament proposat anteriorment de que una comarca pertany a la província on té la majoria de la seva superfície. A més, s'afegeix un camp calculat amb la ràtio de població masculina per cada comarca a l'any 2012, de posterior utilitat per un dels informes requerits. S'afegeixen els corresponents índexs a cadascuna de les taules de dimensió pel guany en velocitat que això suposa. En el cas de la taula de temps el propi any ja fa d'índex.

Així, l'estructura de les taules per les diverses dimensions serà aquesta:



Il·lustració 10: Staging Area

Com es pot apreciar, s'ha creat un índex numèric per a cada dimensió perquè la teoria del model multidimensional de dades així ho recomana. Al crear la taula del fet, les claus textuals es substitueixen pels seus identificadors numèrics i això comporta una millora en la velocitat d'accés per l'aspecte inherent d'optimització que presenten aquest tipus de dades. No s'han creat camps agregats perquè en el seu moment no s'ha considerat necessari en cap cas donades les circumstàncies resultants dels requisits que

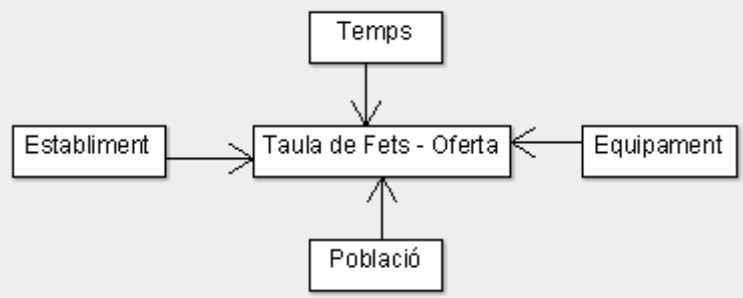
demana el sistema d'intel·ligència de negoci.

Sota la nova estructura de dades que s'ha modificat al curs del projecte s'ha fet una nova estimació del volum de dades que el sistema ha d'emmagatzemar, principalment per la taula dels fets ja que és l'única que s'actualitza i la que concentra la major part de la informació. El volum total de la taula del fet després de la càrrega és d'uns 153.000 registres. Respecte a la mida del registre a emmagatzemar a la taula del Fet es compten 4 bytes per cada camp, que per 9 camps (si s'inclou la clau de la pròpia taula) són 36 bytes per cada registre. Per tant, la mida inicial de la taula és de 153000 * 36 = 5.508.000 bytes = 5,25 Mbytes. Per cada any nou que es carregui a la taula del fet, ja que les magnituds de les diverses dimensions considerades no canviarien a menys que s'introduïssin noves tipologies d'establiments, seria de 22.180 registres nous per cada any, que ocuparien 22180 * 36 = 0,76 Mbytes. Es pot concloure que la nova estimació, tot i haver augmentat considerablement de mida, segueix sent assumible.

3.4. Disseny de la BD – Diagrama E-R

En aquest cas la BD definitiva té la mateixa estructura que la temporal, a excepció d'algun camp que té un format diferent per qüestions d'eficiència espacial. Es va prendre aquesta decisió davant la possibilitat que canviant l'estructura de la BD amb la qual s'havien experimentat pels diversos processos ETL donés problemes al modificar-los per a carregar les dades al esquema definitiu.

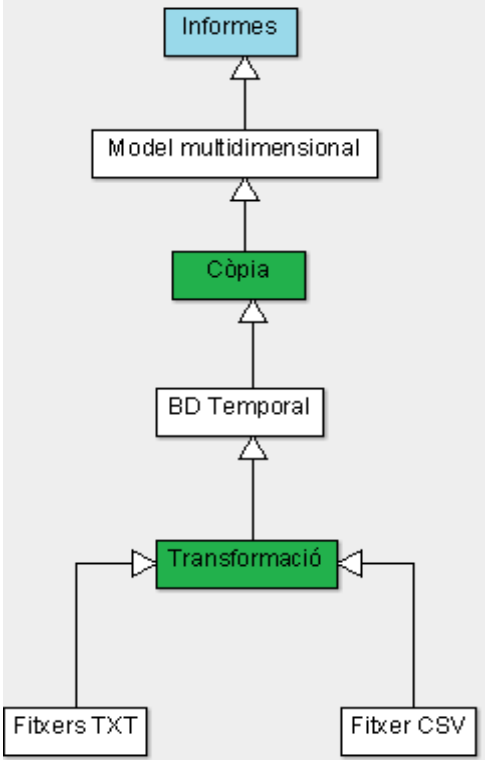
3.4.1. Diagrama E-R



Il·lustració 11: Diagrama E-R

3.5. Procés ETL

Les sigles corresponen a *Extraction, Transformation and Loading*. És a dir, extracció, transformació i càrrega. Es tracta del procés que partint de les fonts de dades origen, arriben a carregar la base de dades destí, fent totes les transformacions que siguin necessàries. En aquest cas, com a programari s'utilitza el que facilita l'empresa Pentaho, PDI o *Pentaho Data Integration*, amb el seu motor *Kettle* i el seu entorn gràfic *Spoon*.
Hi ha diversos fitxers origen, que es volen carregar a unes taules temporals a una base de dades MySQL per fer-hi les posteriors transformacions. Finalment, quan la càrrega d'aquestes dades té èxit a la base de dades temporal, es carreguen les taules que després fan de base pel model multidimensional definitiu. A banda d'aquest procés, cal precisar l'automatització de tot el procés de forma que es pugui fer sense intervenció assistida.



Il·lustració 12: Diagrama del procés ETL

3.6. Errors de Càrrega

Un dels punts importants a qualsevol procés de càrrega és el tractament d'errors, els quals són inevitables quan es tracten amb grans volums de dades. Al detectar qualsevol error, més enllà de la solució que s'empri per la seva resolució o minimització, cal avisar al client per tal que gradualment es vagin netejant les dades. En alguns casos serveix per confirmar si les suposicions inicials són correctes.

3.6.1. Establiments

Hi ha una sèrie d'errors en diversos fitxers. D'una banda, els totals d'establiments per província dels càmpings a l'any 2008 estan mal col·locats. Aquest problema es pot ignorar donat que no s'utilitzen aquestes dades a la càrrega i es calculen en temps d'execució com agregació de les dades per comarca. D'altra banda, també hi ha errors al nombre de places per aquests tipus d'establiments:

- Cases de poble independents de l'any 2008.
- Càmpings de 2a i 3a de l'any 2010.

Afortunadament s'han pogut corregir manualment ja que important els fitxers a Excel i fent camps calculats els errors s'han demostrat evidents.

3.6.2. Equipaments

Durant l'anàlisi prèvia no s'han detectat anomalies.

3.6.3. Població

Durant l'anàlisi prèvia no s'han detectat anomalies.

4. Desenvolupament

En aquest apartat s'explica el treball realitzat per tal d'assolir les fites proposades. Explica els passos realitzats pel desenvolupament del projecte. Un cop finalitzada la fase d'anàlisi i disseny s'ha d'implementar el sistema. El procés no es ben bé lineal, doncs a mesura que es treballa es veu que fa falta refinar o modificar un pas anterior. De forma que el model original es va modificant amb l'objectiu de millorar-lo.

4.1. Programari Utilitzat

Aquesta implementació es realitza sobre una màquina virtual amb Windows XP, servidor de base de dades MySQL i la suite de Pentaho preinstal·lada.
Primer es realitza el procés ETL per incorporar les diferents fonts de dades al sistema. Aquesta tasca es realitza principalment amb l'eina Pentaho Data Integration (PDI o Kettle). Després s'implementa el Cub amb el qual es treballa, amb les dimensions i mesures necessàries, així com altres parametrizacions al sistema. Per aquesta tasca es fa servir el Schema Workbench, que també pertany a la suite Pentaho. Finalment, es fa front a la creació d'informes usant el Pentaho Report Designer (PRD).

Programa	Propòsit
Oracle VM VirtualBox	Executar la màquina virtual
MV Windows XP	Màquina virtual amb sistema operatiu Windows XP
MySQL	Sistema gestor de bases de dades (SGBD)
LibreOffice	Per manipular els fitxers enviats pel client
Microsoft Project 2013	Planificació del projecte i diagrama de Gantt
ArgoUML	Elaboració de diagrames, casos d'ús, E-R, UML
Pentaho Data Integration	Procés ETL
Schema Workbench	Creació del Cub
Pentaho Report Designer	Elaboració dels informes estàtics
Mozilla Firefox	Per accedir al portal i la consola d'administració
Tomcat	Servidor Web
SMRecorder	Per realitzar la presentació

4.2. Procés ETL

La tasca consisteix a integrar la informació distribuïda pel client en diferents formats al sistema fent les transformacions necessàries.

4.2.1. Preparació dels Fitxers

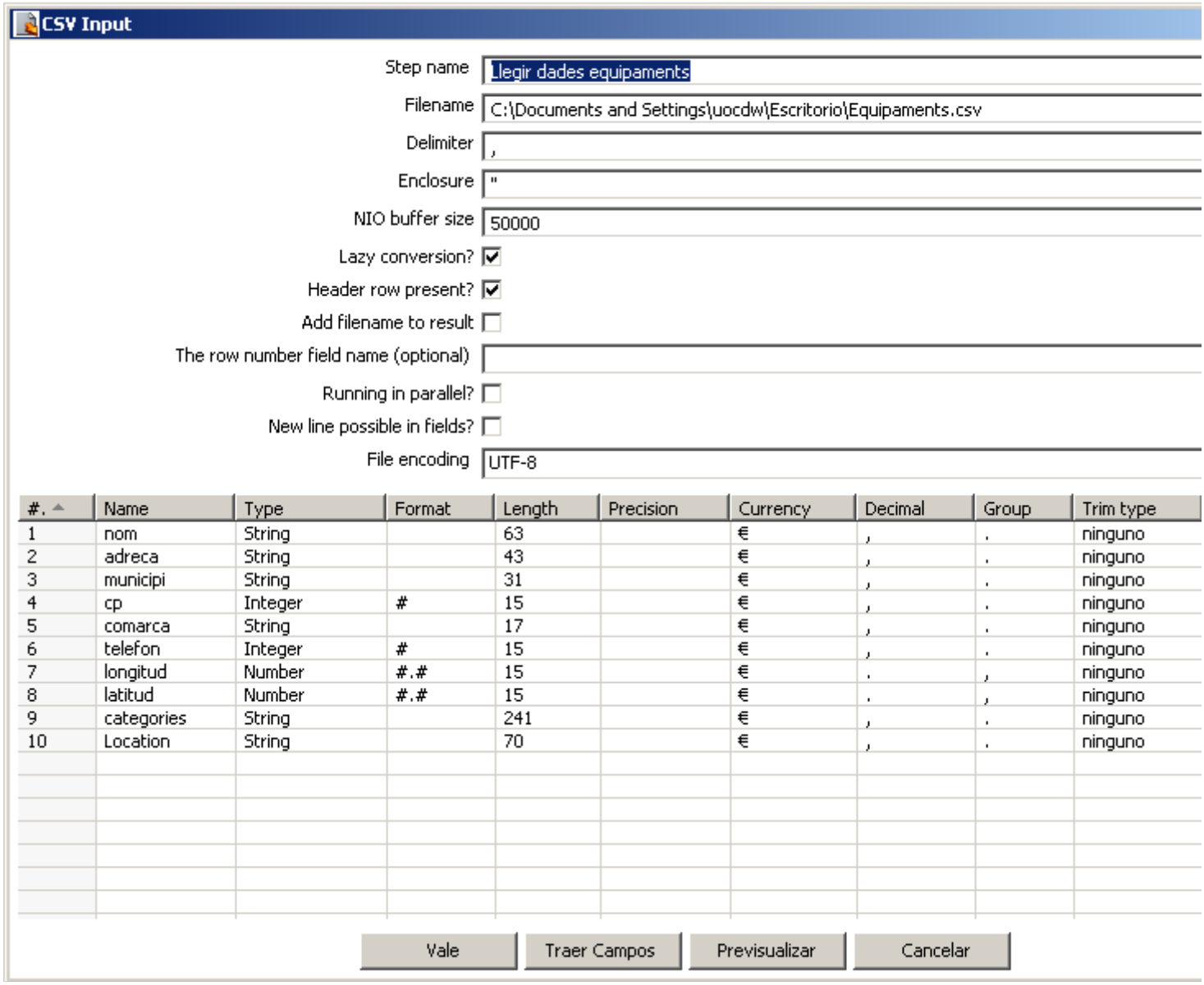
Únicament s'han importat els fitxers d'establiments dels anys 2008 i 2010 a LibreOffice per corregir les dades que tenien errors d'aquests dos fitxers, ja que la resta del procés s'ha dut a terme amb PDI.

4.2.2. Càrrega temporal

En aquest punt es vol passar tota la informació heterogènia a la base de dades MySQL. Per això s'utilitza el programari *Pentaho Data Integration*. El programa permet recuperar la informació de fitxers en diversos formats, entre ells els fitxers de text o csv. Anomena transformacions a cada procés que es genera per realitzar l'homogeneïtzació de les dades, que entre d'altres operacions poden incloure extracció, transformació o càrrega final. Arribat el moment aquestes transformacions es guarden en format XML. A continuació es detalla el procés que s'ha dut a terme per a carregar les dades a cada dimensió del model.

4.2.2.1. Equipaments

En primer lloc s'obre el fitxer d'equipament en format text i es dona format als registres pels diversos camps que hi ha inclosos.



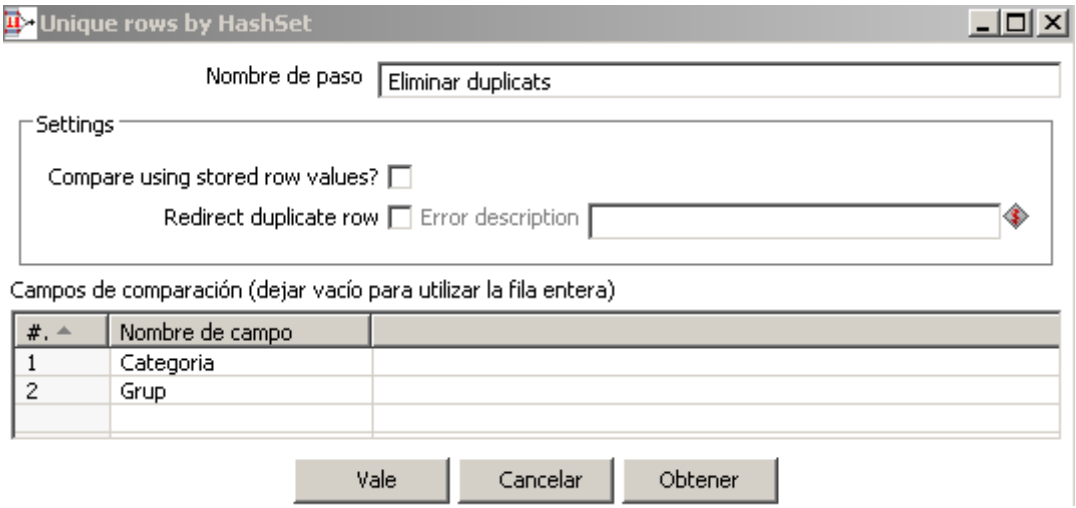
#.	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	nom	String		63		€	,	.	ninguno
2	adreca	String		43		€	,	.	ninguno
3	municipi	String		31		€	,	.	ninguno
4	cp	Integer	#	15		€	,	.	ninguno
5	comarca	String		17		€	,	.	ninguno
6	telefon	Integer	#	15		€	,	.	ninguno
7	longitud	Number	#.#	15		€	.	,	ninguno
8	latitud	Number	#.#	15		€	.	,	ninguno
9	categories	String		241		€	,	.	ninguno
10	Location	String		70		€	,	.	ninguno

Il·lustració 13: Llegir les dades del fitxer d'equipaments

A continuació es selecciona el camp de Categories i es separa per obtenir el grup i categoria d'establiment.

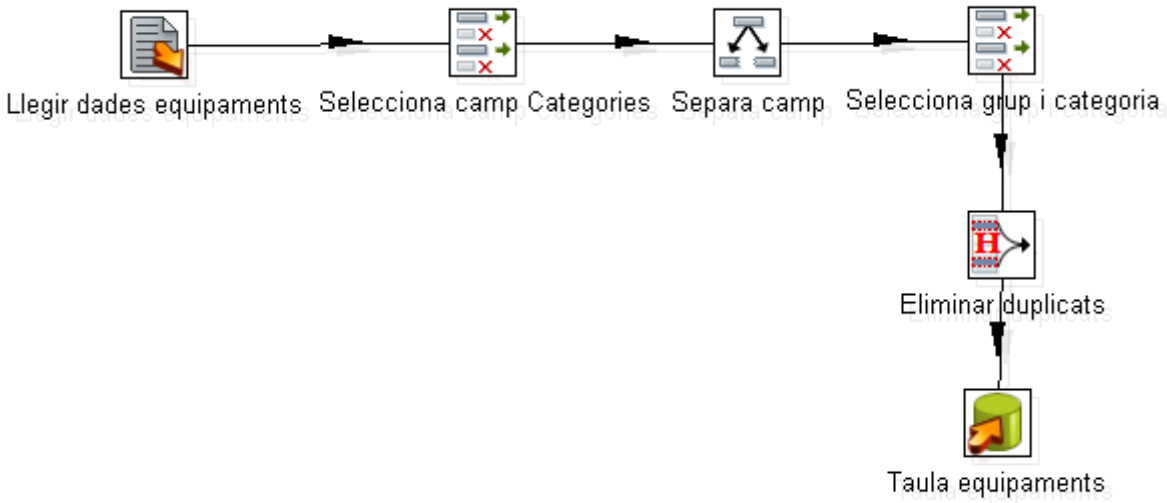
Il·lustració 14: Operador per separar camps de grup i categoria d'equipaments

S'ignora el camp Arrel i s'eliminen els duplicats de la parella Categoria-Grup.



Il·lustració 15: Operador per eliminar duplicats

Finalment, es guarden els registres d'aquest últim pas a la taula d'equipaments. A continuació es mostra el diagrama d'accions que s'han dut a terme per aquesta càrrega.



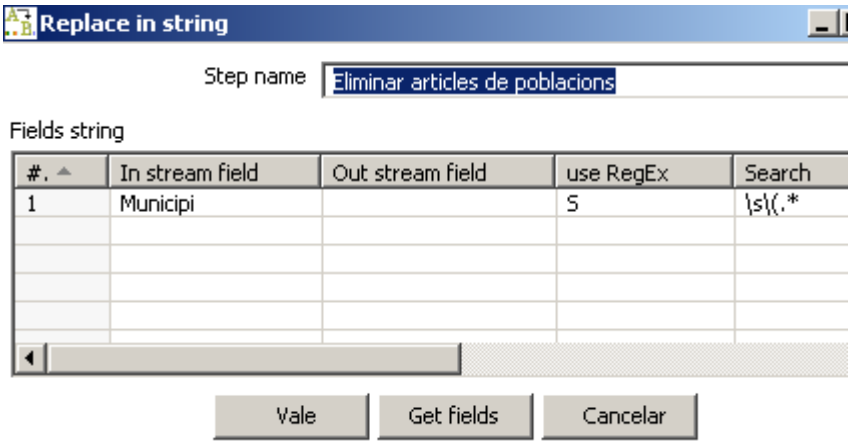
Il·lustració 16: Diagrama ETL Equipaments

4.2.2.2. Població

Primer s'obren els fitxers de població i equipaments. El fitxer d'equipament serveix per obtenir la comarca d'un municipi donat. Fins aquest punt es porten a terme dos processos paral·lels:

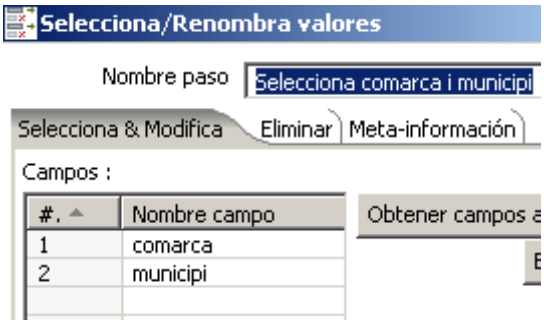
- D'una banda, es seleccionen els camps necessaris del fitxer de poblacions i s'eliminen els articles d'aquelles poblacions que en tinguin mitjançant una expressió regular.

Il·lustració 17: Selecciona camps del fitxer de poblacions

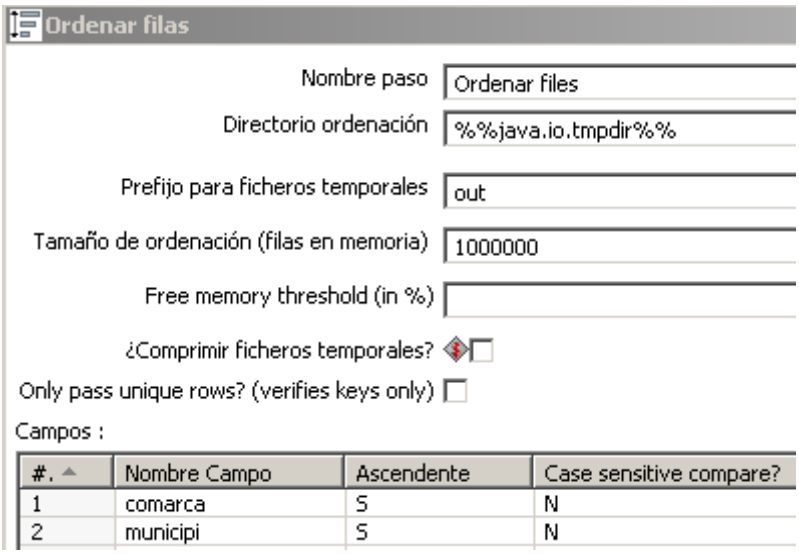


Il·lustració 18: Eliminar articles de les poblacions

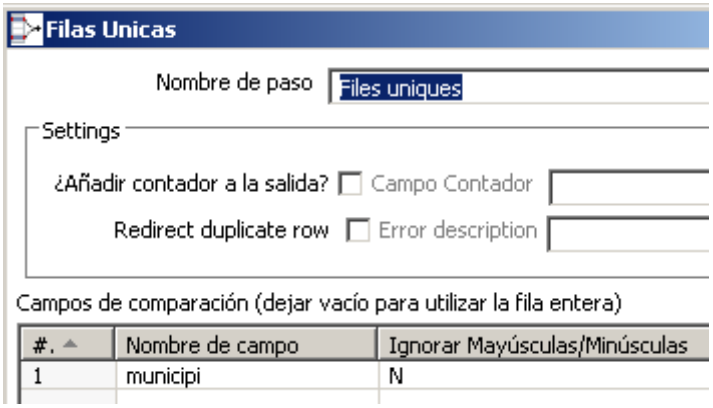
- D'altra banda, es seleccionen la comarca i el municipi del fitxer d'equipaments. Posteriorment, s'eliminen duplicats i articles.



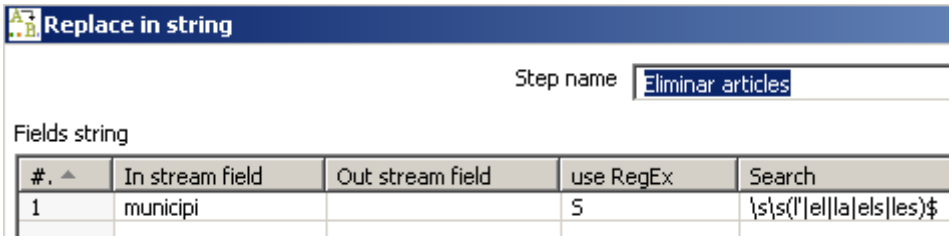
Il·lustració 19: Selecciona comarca i municipi del fitxer d'equipaments



Il·lustració 20: Ordena files per comarca i municipi

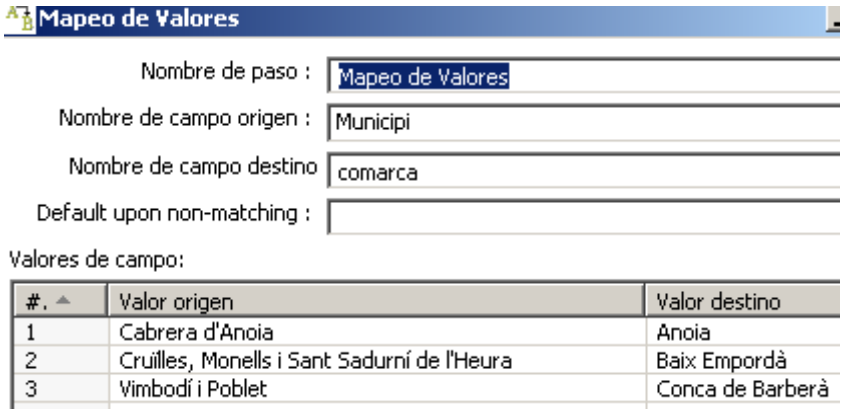


Il·lustració 21: Elimina duplicats



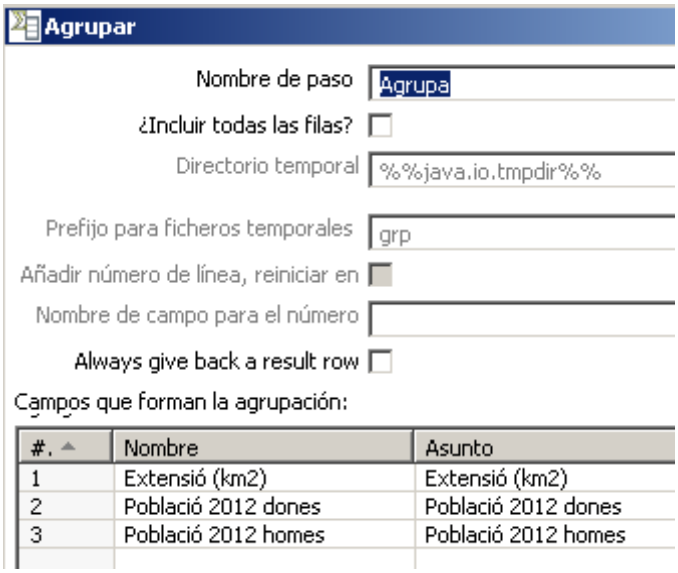
Il·lustració 22: Elimina articles dels municipis del fitxer d'equipaments

Amb les dades del procés pel fitxer d'equipaments s'unifica el flux afegint la comarca al fitxer de poblacions. Posteriorment es filtra ja que hi ha poblacions que per qüestions de nomenclatura no tenen coincidència en els seus noms, pels quals s'afegeix la comarca a cada registre.



Il·lustració 23: Poblacions amb noms especials

Posteriorment s'afegeix la província amb un nou mapa de valors i s'unifiquen els dos fluxos generats pel filtre. S'ordena el flux per comarca i província i s'agrupen els registres calculant els camps agregats necessaris a la base de dades.

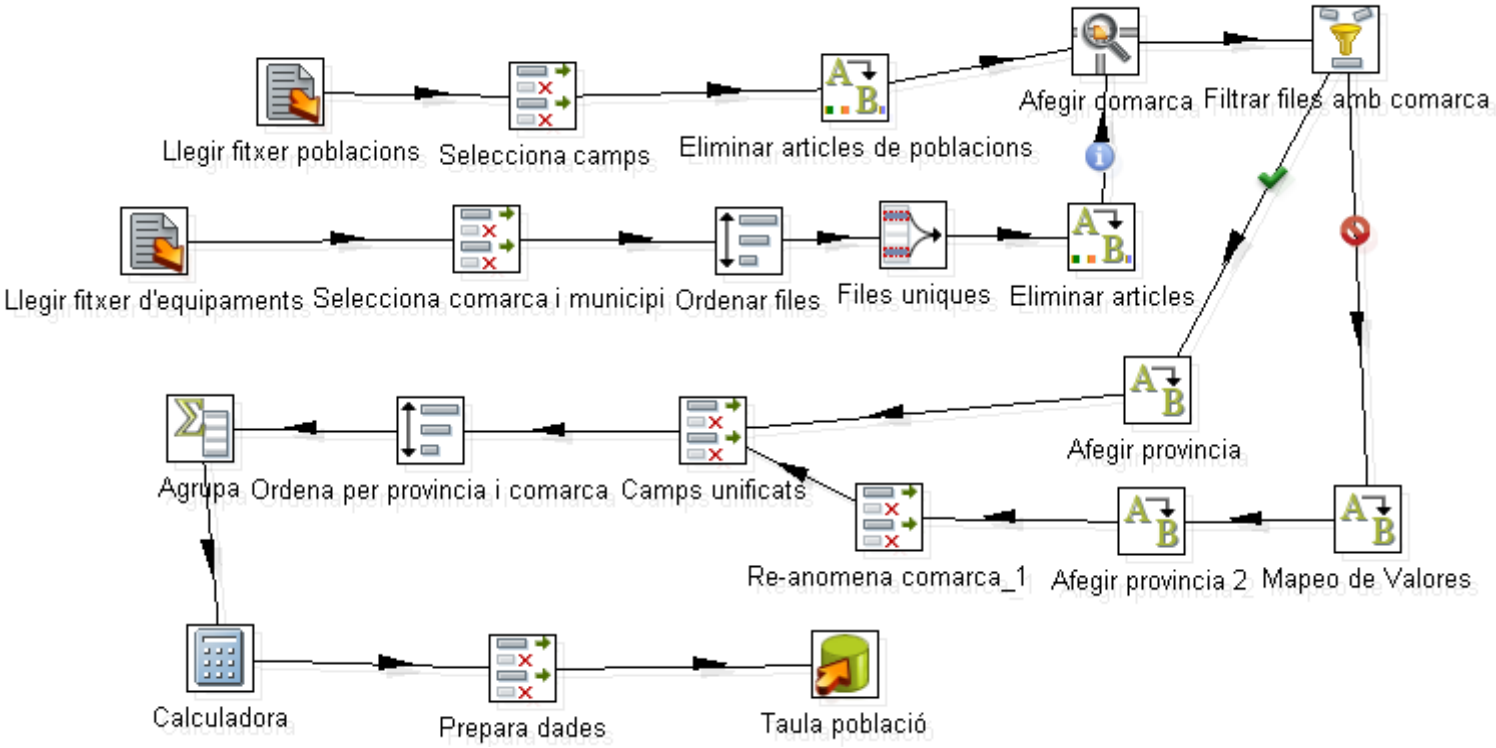


Il·lustració 24: Agrupa i calcula camps agregats per comarca

Després s'utilitza un operador per calcular la ràtio de població masculina per cada comarca.

Il·lustració 25: Calcula nou camp ràtio

Finalment s'introdueixen aquestes dades a la taula de població. A continuació es mostra el diagrama per aquesta transformació.

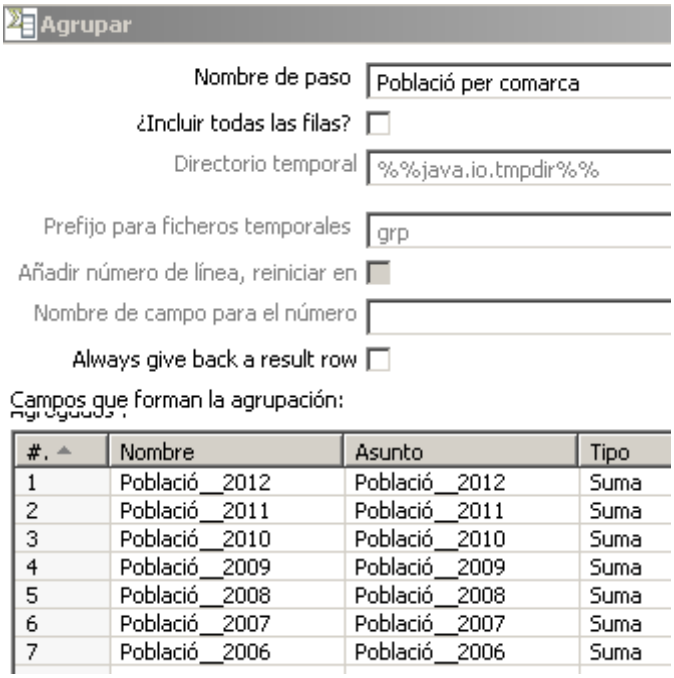


Il·lustració 26: Diagrama ETL Poblacions

4.2.2.3. Oferta

El procés ETL per aquesta taula és bastant més complex que els de les taules anteriors, ja que s'han d'afegir totes les dades útils per les consultes que té el sistema, exceptuant la ràtio i la superfície a la taula de poblacions. Per la complexitat inherent a aquest procés es mostra el procés que es segueix al actualitzar les dades d'oferta per un únic any i després s'expliquen els canvis que hi ha al procés d'automatització de la càrrega per un nou any.

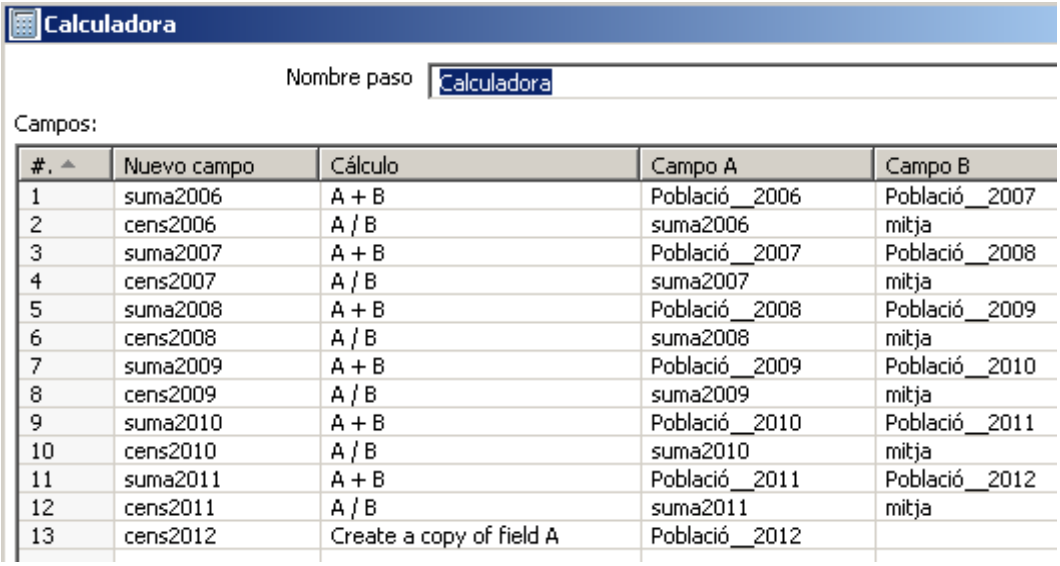
El flux s'inicia als fitxer d'equipaments. Es divideix en dos fluxos separats. D'una banda, es segueix el mateix procés per afegir la província al fitxer de poblacions. Amb aquestes dades a nivell de població, s'agrupen per obtenir els totals per comarca i any.



#.	Nombre	Asunto	Tipo
1	Població__2012	Població__2012	Suma
2	Població__2011	Població__2011	Suma
3	Població__2010	Població__2010	Suma
4	Població__2009	Població__2009	Suma
5	Població__2008	Població__2008	Suma
6	Població__2007	Població__2007	Suma
7	Població__2006	Població__2006	Suma

Il·lustració 27: Agrupació de censos per comarca

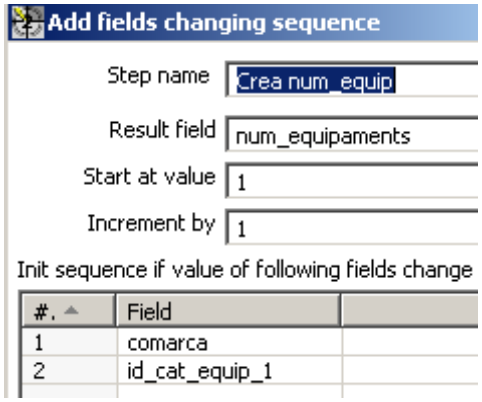
A continuació s'afegeix una constant d'utilitat posterior i es calculen les mitjanes de poblacions a partir de les dades agrupades per comarca entre un any i el següent, a excepció de l'últim any pel qual es fa una còpia.



#.	Nuevo campo	Cálculo	Campo A	Campo B
1	suma2006	A + B	Població__2006	Població__2007
2	cens2006	A / B	suma2006	mitja
3	suma2007	A + B	Població__2007	Població__2008
4	cens2007	A / B	suma2007	mitja
5	suma2008	A + B	Població__2008	Població__2009
6	cens2008	A / B	suma2008	mitja
7	suma2009	A + B	Població__2009	Població__2010
8	cens2009	A / B	suma2009	mitja
9	suma2010	A + B	Població__2010	Població__2011
10	cens2010	A / B	suma2010	mitja
11	suma2011	A + B	Població__2011	Població__2012
12	cens2011	A / B	suma2011	mitja
13	cens2012	Create a copy of field A	Població__2012	

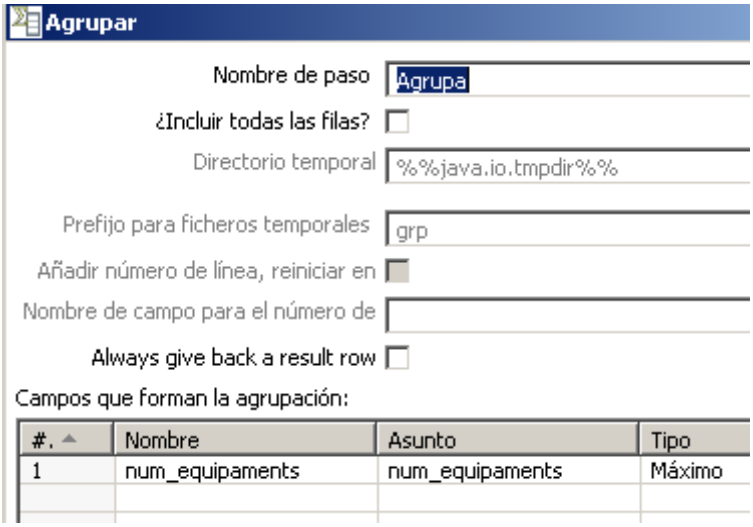
Il·lustració 28: Mitjanes poblacionals

L'altre flux que s'inicia al fitxer d'equipaments és d'utilitat per afegir el nombre d'equipaments per cada comarca. Es seleccionen la comarca i les categories del fitxer d'equipaments. A continuació, es separa el camp de categories per obtenir el grup i la categoria específica. Posteriorment, es fa una consulta a la taula d'equipament per obtenir l'identificador numèric de l'equipament. S'ordenen els registres d'aquest flux i s'aplica un operador que incrementa una variable per cada aparició del equipament en qüestió per crear el camp num_equipaments.



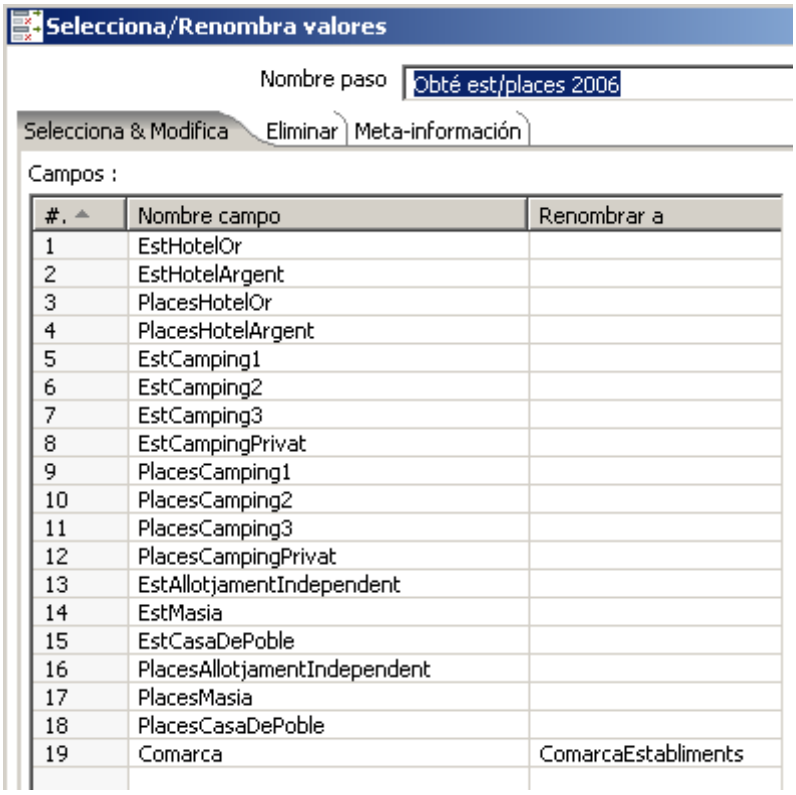
Il·lustració 29: Creació del nombre d'equipaments

Finalment, s'agrupen els resultats guardant únicament l'aparició màxima per cada id_cat_equip_1.



Il·lustració 30: Reducció en el nombre de registres per num_equipament

A partir d'aquí ja es poden crear els registres per cada any considerat. En primer lloc, s'obre el fitxer d'establiment de l'any donat i es recullen els camps d'utilitat.



#.	Nombre campo	Renombrar a
1	EstHotelOr	
2	EstHotelArgent	
3	PlacesHotelOr	
4	PlacesHotelArgent	
5	EstCamping1	
6	EstCamping2	
7	EstCamping3	
8	EstCampingPrivat	
9	PlacesCamping1	
10	PlacesCamping2	
11	PlacesCamping3	
12	PlacesCampingPrivat	
13	EstAllotjamentIndependent	
14	EstMasia	
15	EstCasaDePoble	
16	PlacesAllotjamentIndependent	
17	PlacesMasia	
18	PlacesCasaDePoble	
19	Comarca	ComarcaEstabliments

Il·lustració 31: Camps dels fitxers d'establiments

A continuació, es crea un registre per cada tipus d'establiment amb un operador per normalitzar les files.

#.	Nombre campo	Tipo	campo nuevo
1	EstHotelOr	1	num_establiments
2	EstHotelArgent	2	num_establiments
3	PlacesHotelOr	1	num_places
4	PlacesHotelArgent	2	num_places
5	EstCamping1	5	num_establiments
6	EstCamping2	6	num_establiments
7	EstCamping3	7	num_establiments
8	EstCampingPrivat	8	num_establiments
9	PlacesCamping1	5	num_places
10	PlacesCamping2	6	num_places
11	PlacesCamping3	7	num_places
12	PlacesCampingPrivat	8	num_places
13	EstAllotjamentIndependent	9	num_establiments
14	EstMasia	10	num_establiments
15	EstCasaDePoble	11	num_establiments
16	PlacesAllotjamentIndependent	9	num_places
17	PlacesMasia	10	num_places
18	PlacesCasaDePoble	11	num_places

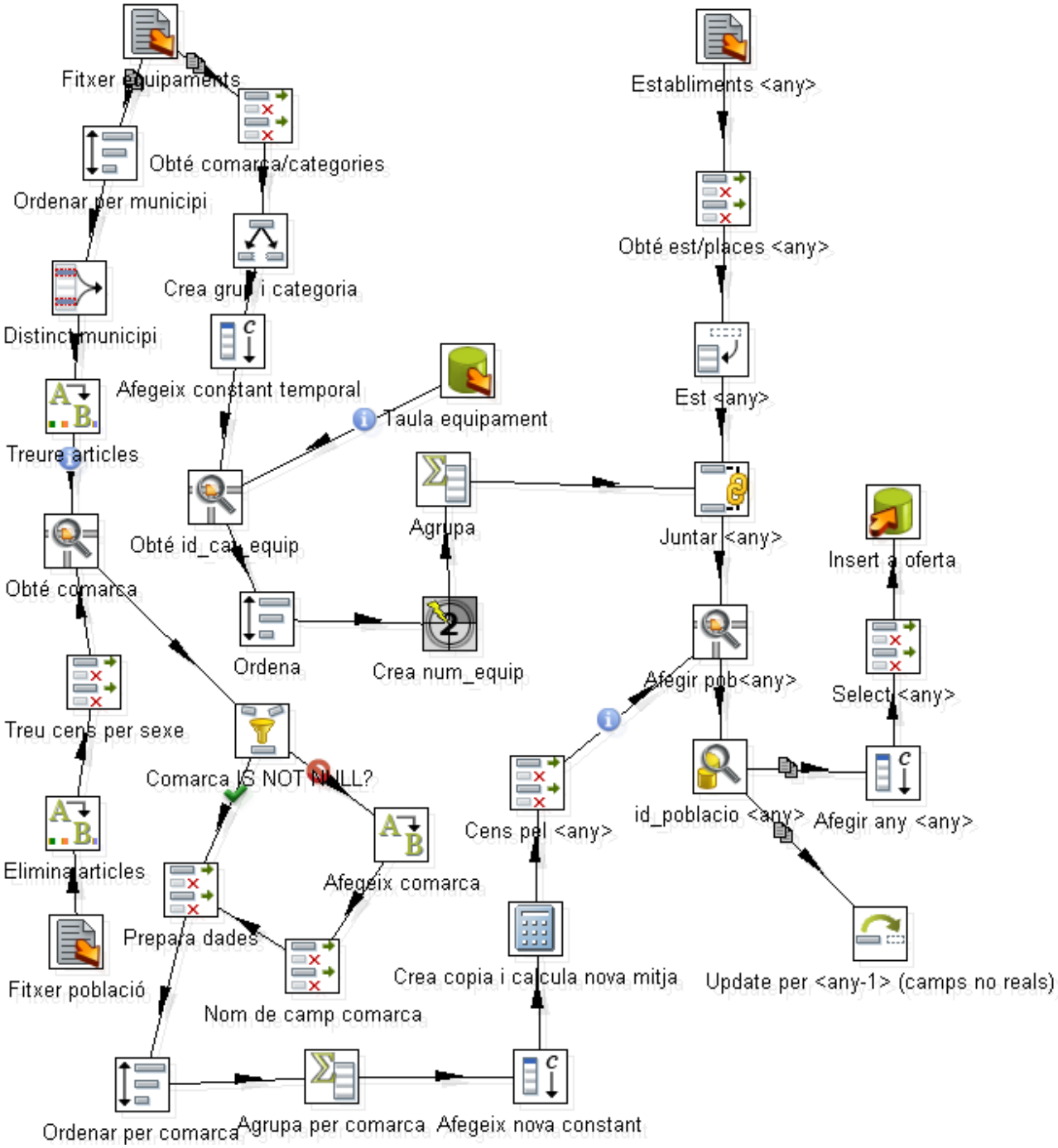
Il·lustració 32: Normalitza files

Posteriorment, s'afegeix el camp num_equipaments procedent del segon flux que es va crear pel fitxer equipaments. Després s'afegeixen les dades del cens per cada comarca procedents del primer flux que es va crear pel fitxer d'equipaments i es canvia la població pel seu codi identificador. Finalment, es seleccionen i s'unifiquen els camps necessaris per cada any, s'agrupen els registres per tots els anys i s'inserta a la taula dels fets.

La diferència principal que té el procés d'automatització per les noves dades és que, després d'obtenir el codi identificador per cada comarca, el flux es divideix. D'una banda, es segueix el procés que té lloc a la càrrega inicial, en aquest cas inserint els registres pel nou any a la taula dels fets. En aquest cas serien els registres per l'any 2013. Però d'altra banda s'aplica un operador d'actualització pels registres de l'any anterior (que en aquest cas seria 2012) amb el nou valor pel camp de població com la mitjana entre les dades de població entre 2012 i 2013.

Il·lustració 33: Actualització pels registres de l'any anterior

Cal fer notar que els noms dels camps durant tot el procés d'automatització no són reals i només serveixen per il·lustrar els noms reals que han de tenir a l'hora de fer la càrrega de noves dades. El procés que es porta a terme per l'automatització en la càrrega de noves dades, i que guarda moltes semblances al de la càrrega inicial, es mostra a continuació.



Il·lustració 34: Càrrega de noves dades a la taula dels fets

4.2.3. Càrrega definitiva

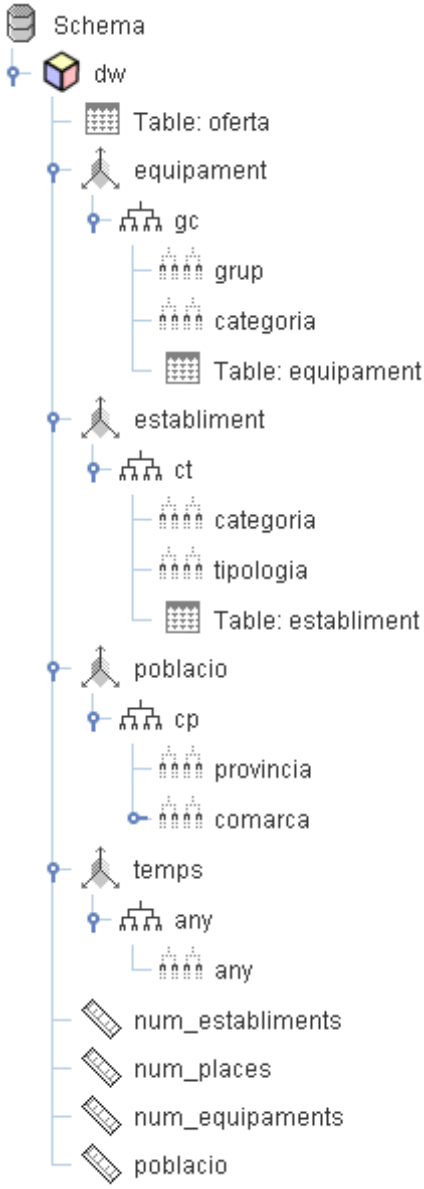
L'únic canvi que es porta a terme als processos ETL és el canvi de l'esquema de la base de dades de prova (test) al esquema definitiu (DW) a tots aquells operadors que insereix registres a la base de dades, ja que no hi ha canvis substancials en l'estructura d'una base de dades i l'altra.

4.2.4. Observacions

En general, la càrrega de les taules a la base de dades ha anat força bé. L'excepció a aquesta afirmació ha estat la taula dels fets. La gran complexitat del procés ha fet que la majoria d'ocasions no s'insertessin tots els registres, així que a la càrrega per aquesta taula es va optar per dues alternatives. Inicialment, ja que les dades de l'esquema de prova eren correctes, es va fer una còpia directa de tots els registres de l'esquema de test al esquema pel model multidimensional. Posteriorment, es va veure la necessitat de modificar el procés perquè s'havia oblidat la part a la qual els censos de població són la mitjana entre un any i el següent. A més de modificar l'ETL per la taula dels fets, es va decidir fer passar el procés mitjançant l'opció "Preview" perquè s'anessin mostrant tots els registres que el procés anava generant. Cada vegada que es feia clic a l'opció per mostrar més registres, el procés tornava a iterar per generar 1000 nous registres i els mostrava a l'usuari. Tenint en compte que la taula del fet conté més de 153.000 registres es pot concloure que el procés va arribar a ser molt tediós.

4.3. Model Multidimensional

Fent servir l'eina *Schema Workbench* es comença a dissenyar el cub amb el qual es treballa. El cub és diu "dw". En aquest esquema s'indiquen les diferents dimensions, amb la taula que es recolzen. A cada dimensió se li assigna una o més jerarquies, amb els seus diferents nivells. Per últim s'assignen al cub.



Il·lustració 35: Esquema del cub

Un cop finalitzat la modelització es puja al servidor per poder utilitzar-lo des del portal. Per poder realitzar aquesta tasca (i després de forma anàloga en la publicació dels informes) cal editar un fitxer de configuració per informar la contrasenya que es fa servir per tal de poder publicar el cub. Un cop al servidor s'emmagatzema el cub que s'ha creat.

4.4. Consultes

Fins aquí la major part de tasques han estat transparents pel usuari. Ara cal desenvolupar els diferents informes que s'han sol·licitat.
Primer cal crear un entorn de treball, que en aquest cas particular serà Pentaho Report Designer, per les capacitats afegides que ofereix respecte a l'eina per generar informes des del portal. Després, amb els informes publicats des de Pentaho Report Designer, es pot accedir al portal pel port 8080 i utilitzar-los de forma transparent pel usuari.

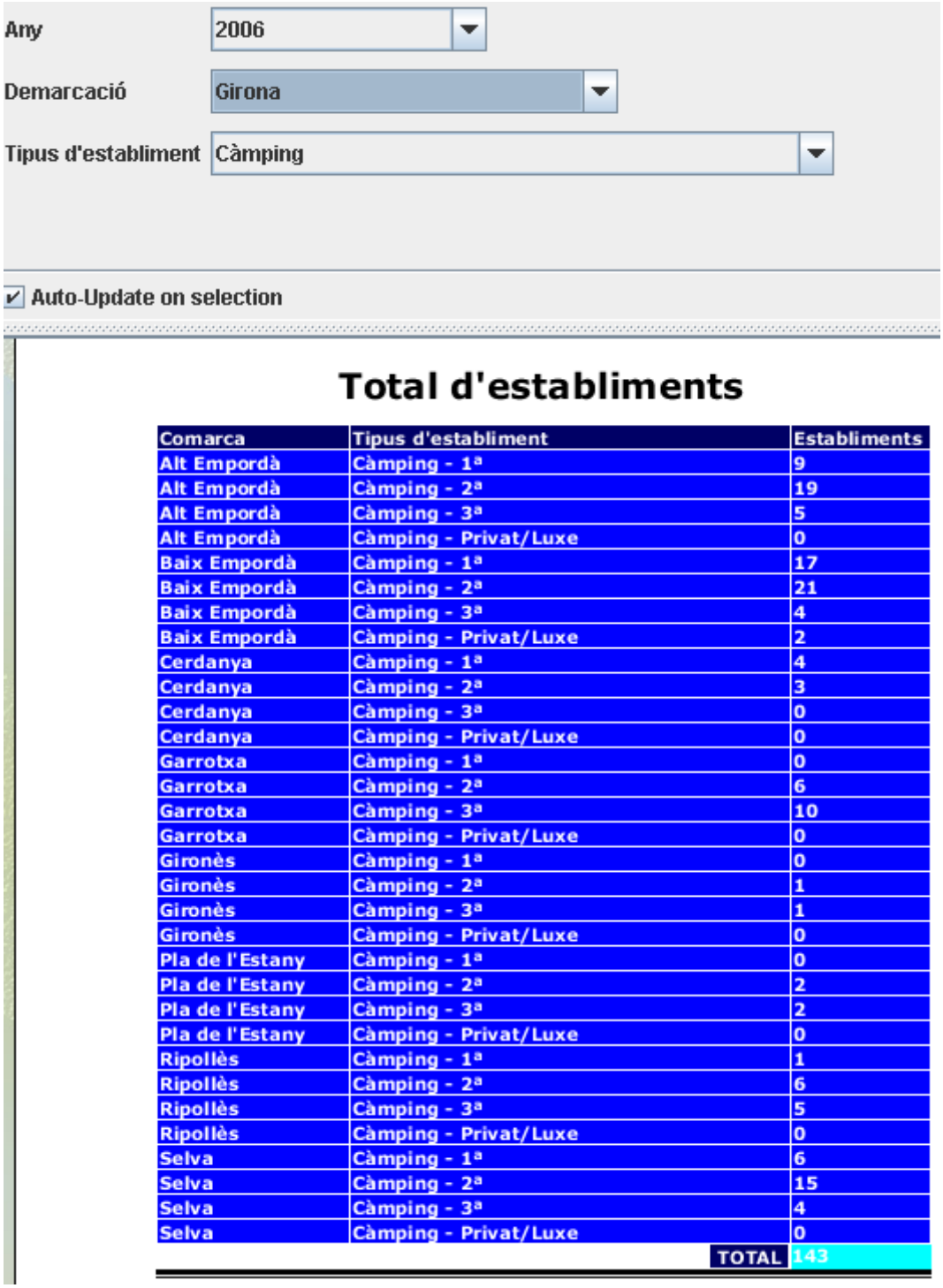
Per fer ús del portal i les seves eines s'afegeix un nou *Datasource*, amb el qual s'indica com connectar amb la base de dades. S'afegeix a la consola d'administració, a la qual es pot accedir des del port 8099, i després s'afegeix també al fitxer de configuració *default.properties* al directori *sample-jndi.*

4.4.1. Informes i Anàlisi

S'han tingut en compte els diversos canvis en el tipus d'establiments disponibles al llargs dels anys a partir de l'any que seleccioni el usuari. Per exemple, si el usuari escolleix l'any 2006 no podrà escollir una masoveria com el tipus d'establiment que desitja consultar. A més, s'ha harmonitzat el nombre de places de càmpings pel canvi en el mètode de càlcul que va haver-hi entre 2010 i 2011.
S'han afegit camps totals al peu de l'informe en cas que l'usuari seleccioni un grup d'establiments, per exemples tots els establiments hotelers, o una província que és resultat de l'agregació pels registres per comarca.
Tots els informes admeten paràmetres per tal de seleccionar l'any de consulta, la demarcació escollida i la categoria en aquells informes on sigui necessari. En el cas de l'oferta mitjana de places es poden escollir l'any inicial i l'any final pel qual es desitja calcular la oferta mitjana. Aquesta decisió es deu al rendiment de la màquina virtual, per tal de fer ús de les facilitats d'*slicing* del model multidimensional i obtenir un rendiment millor. L'*slicing* consisteix en seccionar part del cub per tan sols haver de treballar amb un tros. Amb el model relacional té molta semblança amb l'ús de les clàusules *WHERE*, tant en sintaxi com en resultat. Sobre aquest tema cal destacar que Mondrian, el motor ROLAP que es fa servir, fa ús d'una memòria cau que provoca que el primer cop que s'executa la consulta és quan triga més.

1. Total d'establiments

Per aquesta consulta s'han llistat els nombre d'establiments a partir del tipus d'establiment, la demarcació i l'any sol·licitats



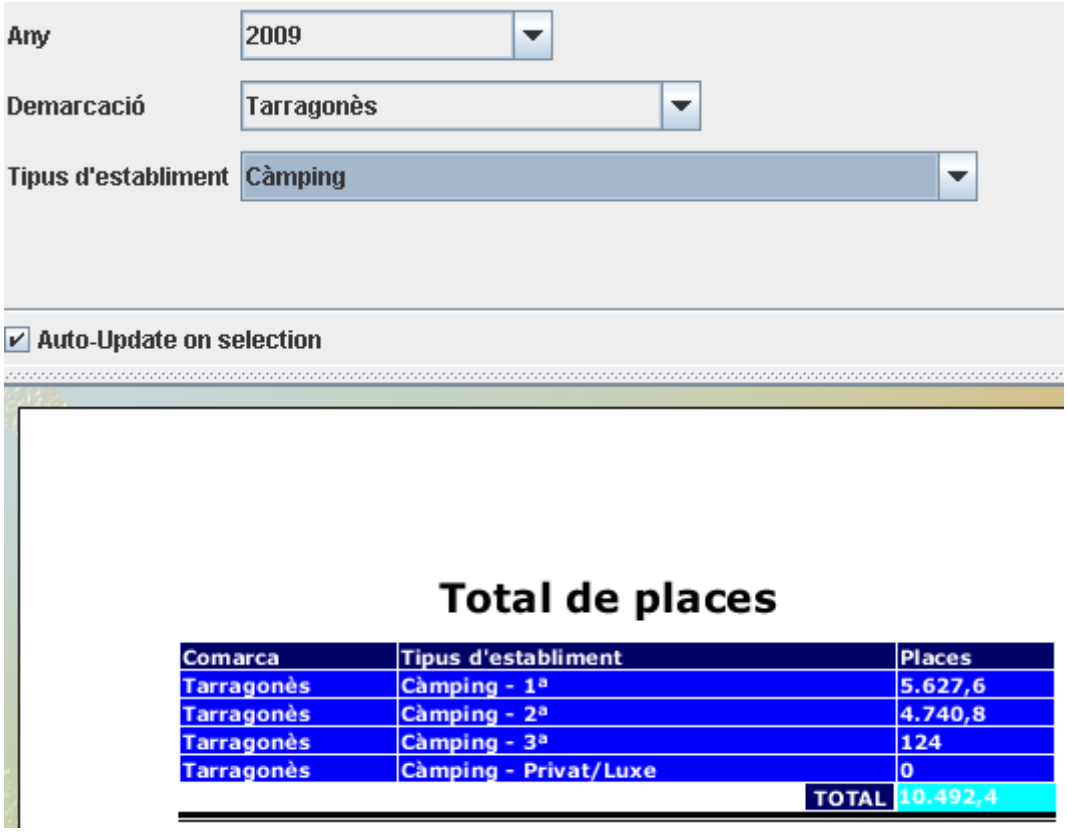
Total d'establiments

Comarca	Tipus d'establiment	Establiments
Alt Empordà	Càmping - 1ª	9
Alt Empordà	Càmping - 2ª	19
Alt Empordà	Càmping - 3ª	5
Alt Empordà	Càmping - Privat/Luxe	0
Baix Empordà	Càmping - 1ª	17
Baix Empordà	Càmping - 2ª	21
Baix Empordà	Càmping - 3ª	4
Baix Empordà	Càmping - Privat/Luxe	2
Cerdanya	Càmping - 1ª	4
Cerdanya	Càmping - 2ª	3
Cerdanya	Càmping - 3ª	0
Cerdanya	Càmping - Privat/Luxe	0
Garrotxa	Càmping - 1ª	0
Garrotxa	Càmping - 2ª	6
Garrotxa	Càmping - 3ª	10
Garrotxa	Càmping - Privat/Luxe	0
Gironès	Càmping - 1ª	0
Gironès	Càmping - 2ª	1
Gironès	Càmping - 3ª	1
Gironès	Càmping - Privat/Luxe	0
Pla de l'Estany	Càmping - 1ª	0
Pla de l'Estany	Càmping - 2ª	2
Pla de l'Estany	Càmping - 3ª	2
Pla de l'Estany	Càmping - Privat/Luxe	0
Ripollès	Càmping - 1ª	1
Ripollès	Càmping - 2ª	6
Ripollès	Càmping - 3ª	5
Ripollès	Càmping - Privat/Luxe	0
Selva	Càmping - 1ª	6
Selva	Càmping - 2ª	15
Selva	Càmping - 3ª	4
Selva	Càmping - Privat/Luxe	0
	TOTAL	143

Il·lustració 36: Informe del total d'establiments

2. Total de places

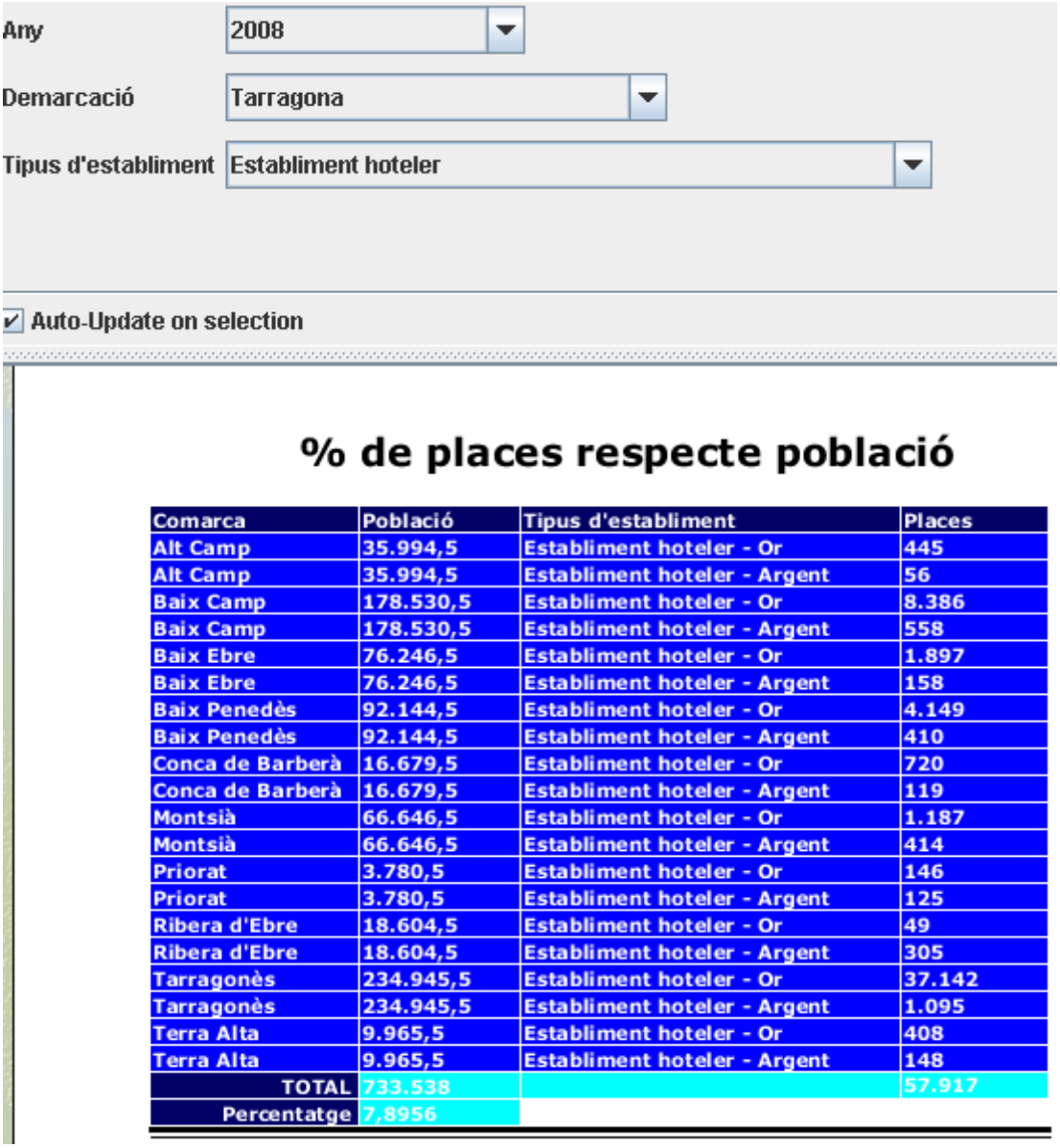
El informe és semblant a l'anterior però canviant el nombre d'establiments pel nombre de places, i tenint en compte la casuística especial del nombre de places pels càmpings.



Il·lustració 37: Informe del total de places

3. % de places respecte població

Es mostra un percentatge del nombre de places que hi ha per una demarcació, tipus d'establiment i any requerit respecte al cens de població de dita demarcació per l'any concret.



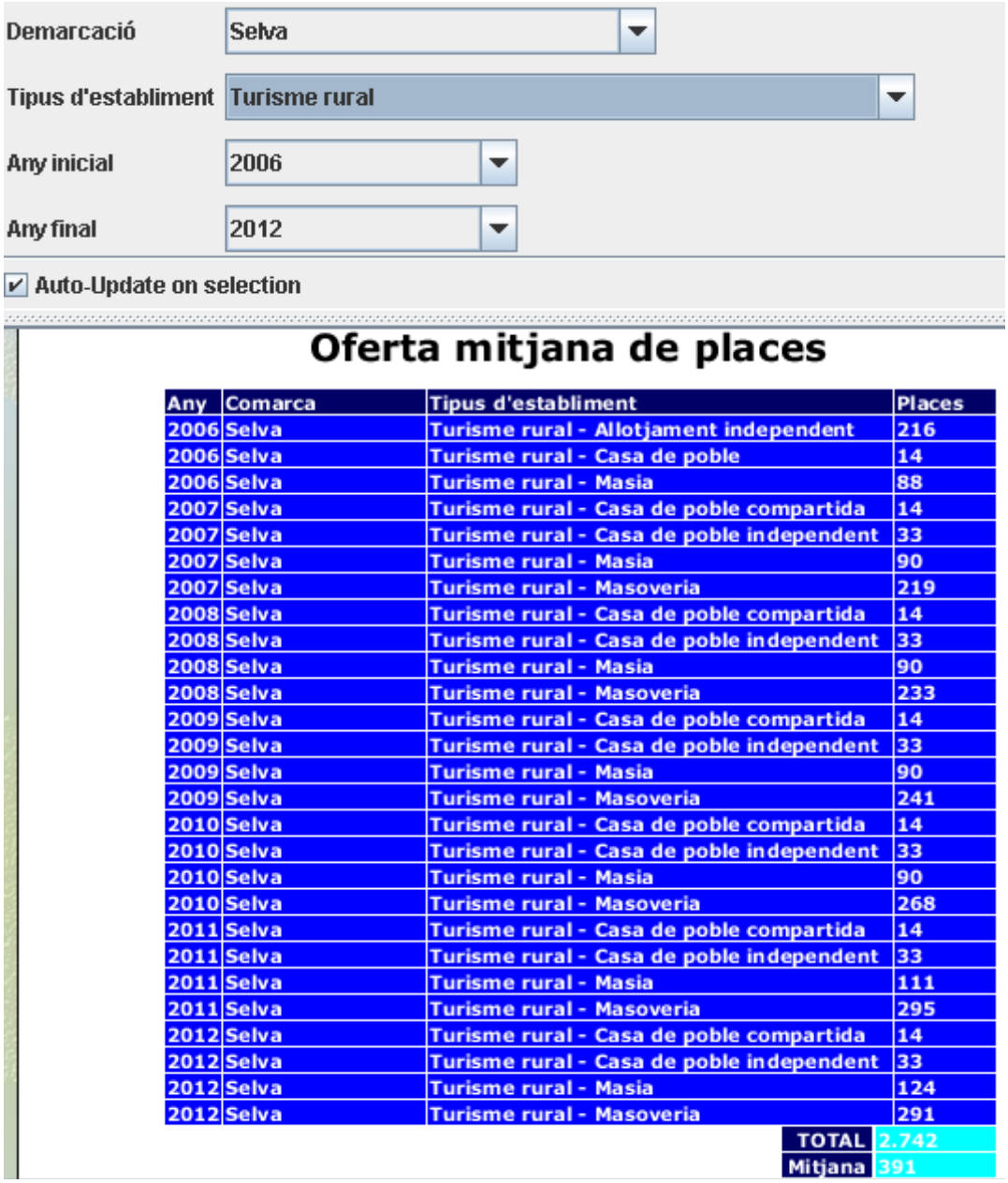
% de places respecte població

Comarca	Població	Tipus d'establiment	Places
Alt Camp	35.994,5	Establiment hoteler - Or	445
Alt Camp	35.994,5	Establiment hoteler - Argent	56
Baix Camp	178.530,5	Establiment hoteler - Or	8.386
Baix Camp	178.530,5	Establiment hoteler - Argent	558
Baix Ebre	76.246,5	Establiment hoteler - Or	1.897
Baix Ebre	76.246,5	Establiment hoteler - Argent	158
Baix Penedès	92.144,5	Establiment hoteler - Or	4.149
Baix Penedès	92.144,5	Establiment hoteler - Argent	410
Conca de Barberà	16.679,5	Establiment hoteler - Or	720
Conca de Barberà	16.679,5	Establiment hoteler - Argent	119
Montsià	66.646,5	Establiment hoteler - Or	1.187
Montsià	66.646,5	Establiment hoteler - Argent	414
Priorat	3.780,5	Establiment hoteler - Or	146
Priorat	3.780,5	Establiment hoteler - Argent	125
Ribera d'Ebre	18.604,5	Establiment hoteler - Or	49
Ribera d'Ebre	18.604,5	Establiment hoteler - Argent	305
Tarragonès	234.945,5	Establiment hoteler - Or	37.142
Tarragonès	234.945,5	Establiment hoteler - Argent	1.095
Terra Alta	9.965,5	Establiment hoteler - Or	408
Terra Alta	9.965,5	Establiment hoteler - Argent	148
TOTAL	733.538		57.917
Percentatge	7,8956		

Il·lustració 38: Informe del percentatge de places respecte la població

4. Oferta mitjana de places

Es mostra el nombre mitjà de places a partir d'un període concret per una demarcació i tipus d'establiment sol·licitat.



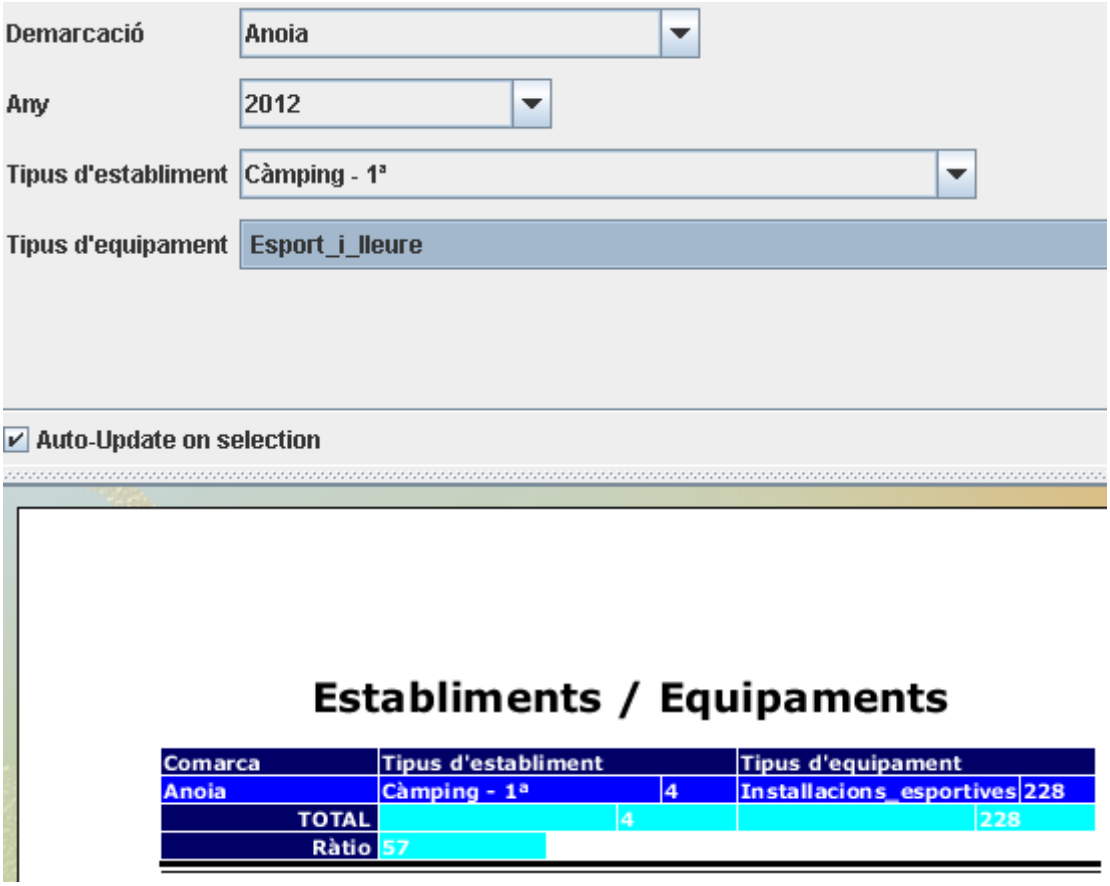
Oferta mitjana de places

Any	Comarca	Tipus d'establiment	Places
2006	Selva	Turisme rural - Allotjament independent	216
2006	Selva	Turisme rural - Casa de poble	14
2006	Selva	Turisme rural - Masia	88
2007	Selva	Turisme rural - Casa de poble compartida	14
2007	Selva	Turisme rural - Casa de poble independent	33
2007	Selva	Turisme rural - Masia	90
2007	Selva	Turisme rural - Masoveria	219
2008	Selva	Turisme rural - Casa de poble compartida	14
2008	Selva	Turisme rural - Casa de poble independent	33
2008	Selva	Turisme rural - Masia	90
2008	Selva	Turisme rural - Masoveria	233
2009	Selva	Turisme rural - Casa de poble compartida	14
2009	Selva	Turisme rural - Casa de poble independent	33
2009	Selva	Turisme rural - Masia	90
2009	Selva	Turisme rural - Masoveria	241
2010	Selva	Turisme rural - Casa de poble compartida	14
2010	Selva	Turisme rural - Casa de poble independent	33
2010	Selva	Turisme rural - Masia	90
2010	Selva	Turisme rural - Masoveria	268
2011	Selva	Turisme rural - Casa de poble compartida	14
2011	Selva	Turisme rural - Casa de poble independent	33
2011	Selva	Turisme rural - Masia	111
2011	Selva	Turisme rural - Masoveria	295
2012	Selva	Turisme rural - Casa de poble compartida	14
2012	Selva	Turisme rural - Casa de poble independent	33
2012	Selva	Turisme rural - Masia	124
2012	Selva	Turisme rural - Masoveria	291
		TOTAL	2.742
		Mitjana	391

Il·lustració 39: Informe de la oferta mitjana de places

5. Nombre d'establiments / Nombre d'equipaments

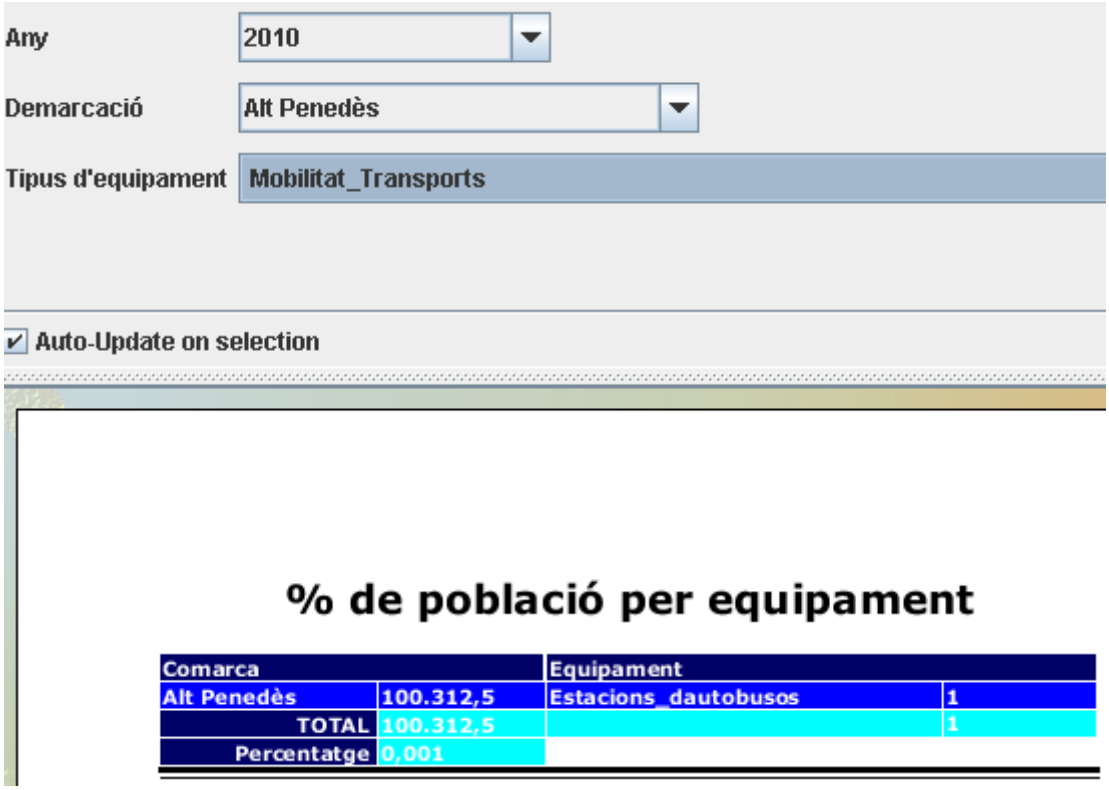
Es visualitza una ràtio del nombre d'equipaments públics que hi ha disponibles respecte cada establiment turístic donat un any, demarcació i tipus d'establiments i d'equipaments sol·licitats.



Il·lustració 40: Informe d'establiments vs equipaments

6. % de població per equipament

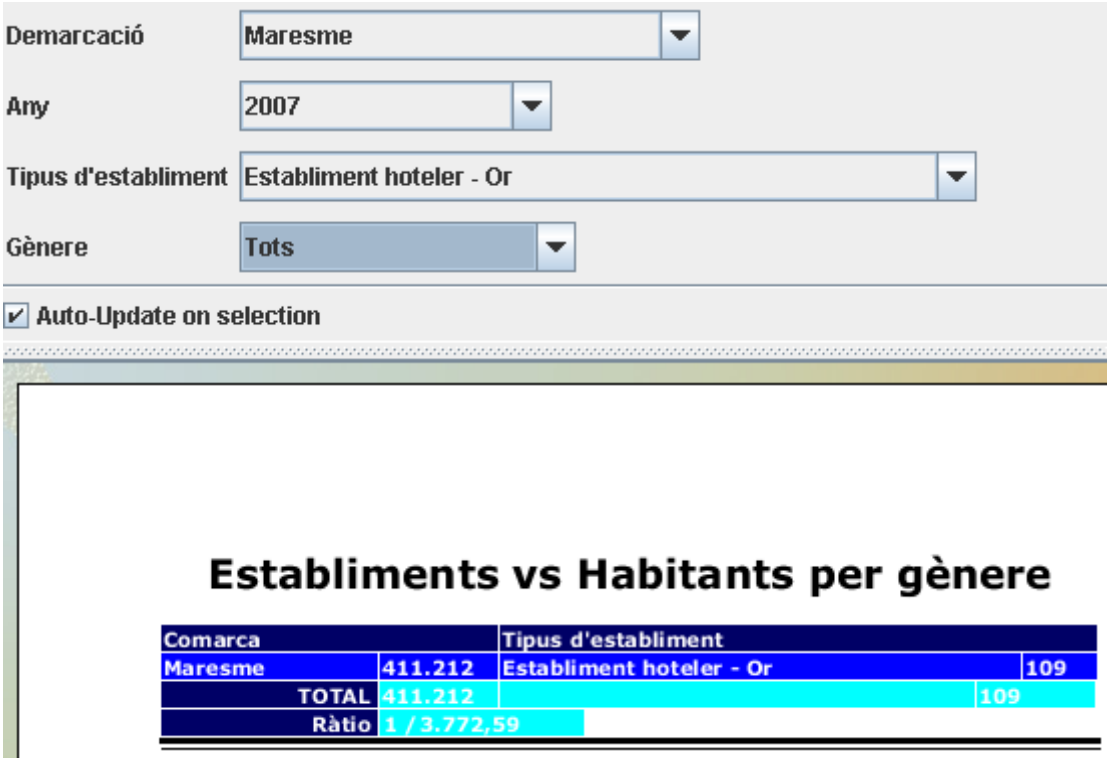
Es mostra un percentatge del nombre d'equipaments públics que hi ha disponibles respecte el cens de població d'una demarcació determinada per un any donat.



Il·lustració 41: Informe del percentatge de població per equipament

7. Indicador d'establiments vs habitants per gènere

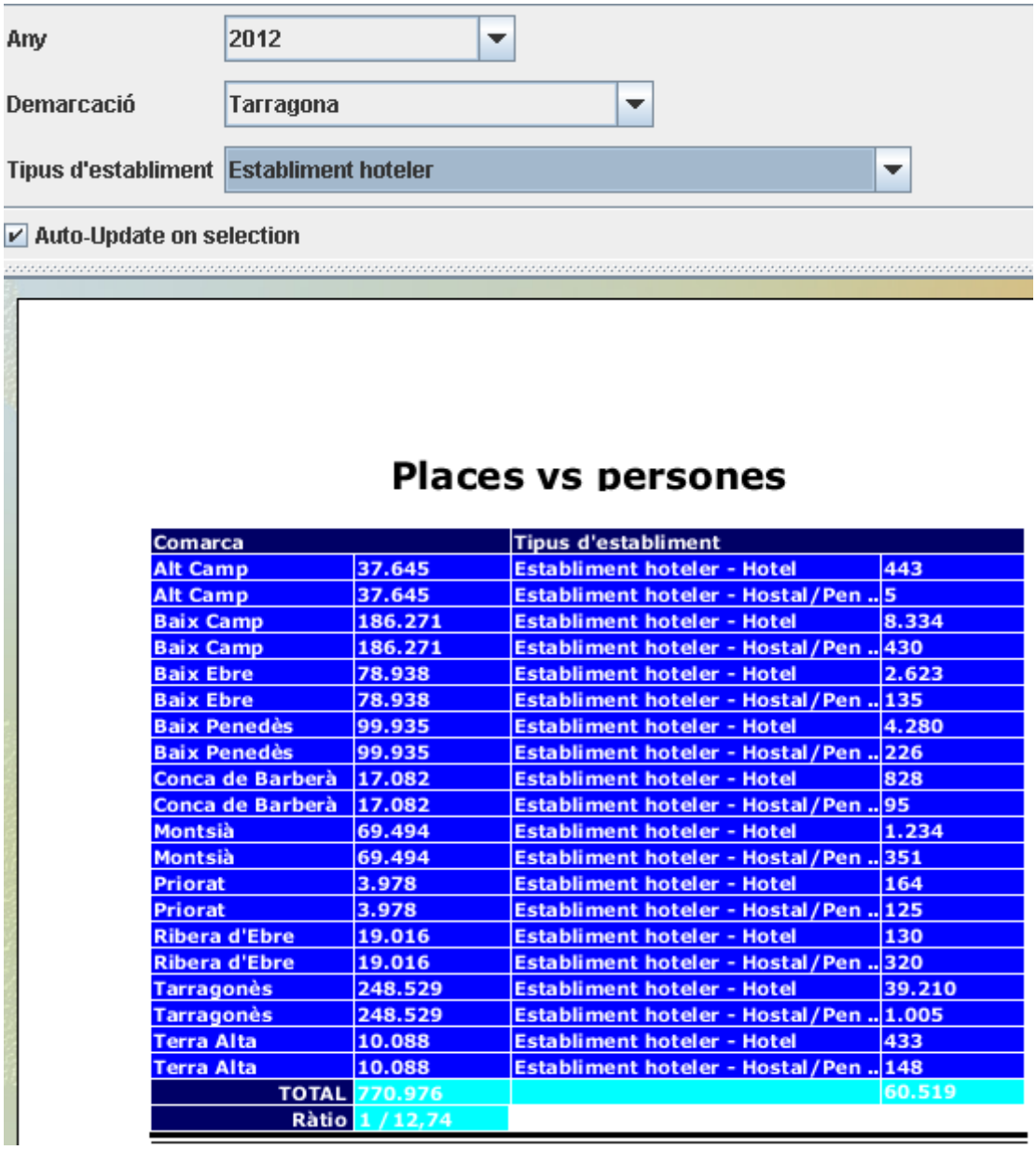
Es mostra una ràtio que representa el nombre d'habitants, seleccionats per gènere si així es desitja, que hi ha per cada establiment turístic d'un tipus determinat, per una demarcació i any sol·licitats.



Il·lustració 42: Informe d'establiments vs habitants per gènere

8. Indicador de places vs persones

Es mostra una ràtio que és la proporció de persones que resideixen en una demarcació determinada per cada establiment turístic d'un tipus determinat per un any donat.



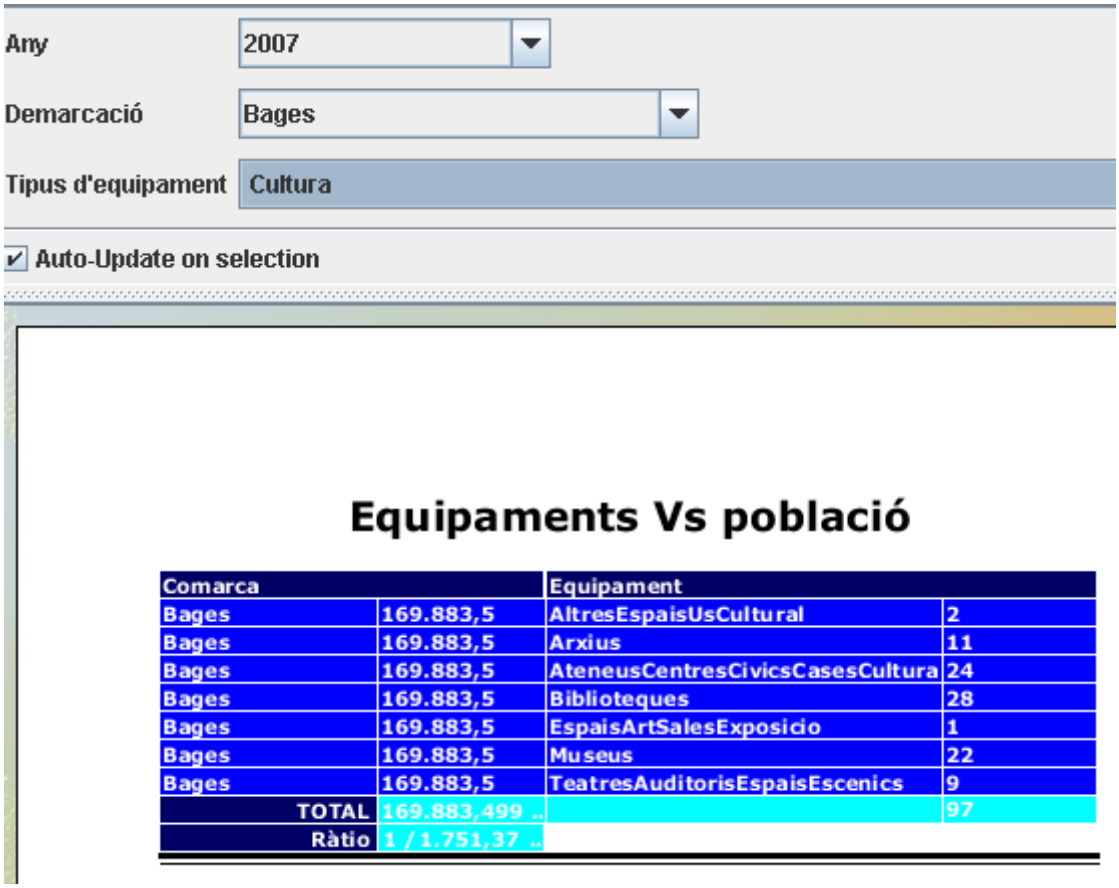
Places vs persones

Comarca		Tipus d'establiment	
Alt Camp	37.645	Establiment hoteler - Hotel	443
Alt Camp	37.645	Establiment hoteler - Hostal/Pen ..	5
Baix Camp	186.271	Establiment hoteler - Hotel	8.334
Baix Camp	186.271	Establiment hoteler - Hostal/Pen ..	430
Baix Ebre	78.938	Establiment hoteler - Hotel	2.623
Baix Ebre	78.938	Establiment hoteler - Hostal/Pen ..	135
Baix Penedès	99.935	Establiment hoteler - Hotel	4.280
Baix Penedès	99.935	Establiment hoteler - Hostal/Pen ..	226
Conca de Barberà	17.082	Establiment hoteler - Hotel	828
Conca de Barberà	17.082	Establiment hoteler - Hostal/Pen ..	95
Montsià	69.494	Establiment hoteler - Hotel	1.234
Montsià	69.494	Establiment hoteler - Hostal/Pen ..	351
Priorat	3.978	Establiment hoteler - Hotel	164
Priorat	3.978	Establiment hoteler - Hostal/Pen ..	125
Ribera d'Ebre	19.016	Establiment hoteler - Hotel	130
Ribera d'Ebre	19.016	Establiment hoteler - Hostal/Pen ..	320
Tarragonès	248.529	Establiment hoteler - Hotel	39.210
Tarragonès	248.529	Establiment hoteler - Hostal/Pen ..	1.005
Terra Alta	10.088	Establiment hoteler - Hotel	433
Terra Alta	10.088	Establiment hoteler - Hostal/Pen ..	148
TOTAL	770.976		60.519
Ràtio	1 / 12,74		

Il·lustració 43: Informe de places vs persones

9. Indicador d'equipaments vs població

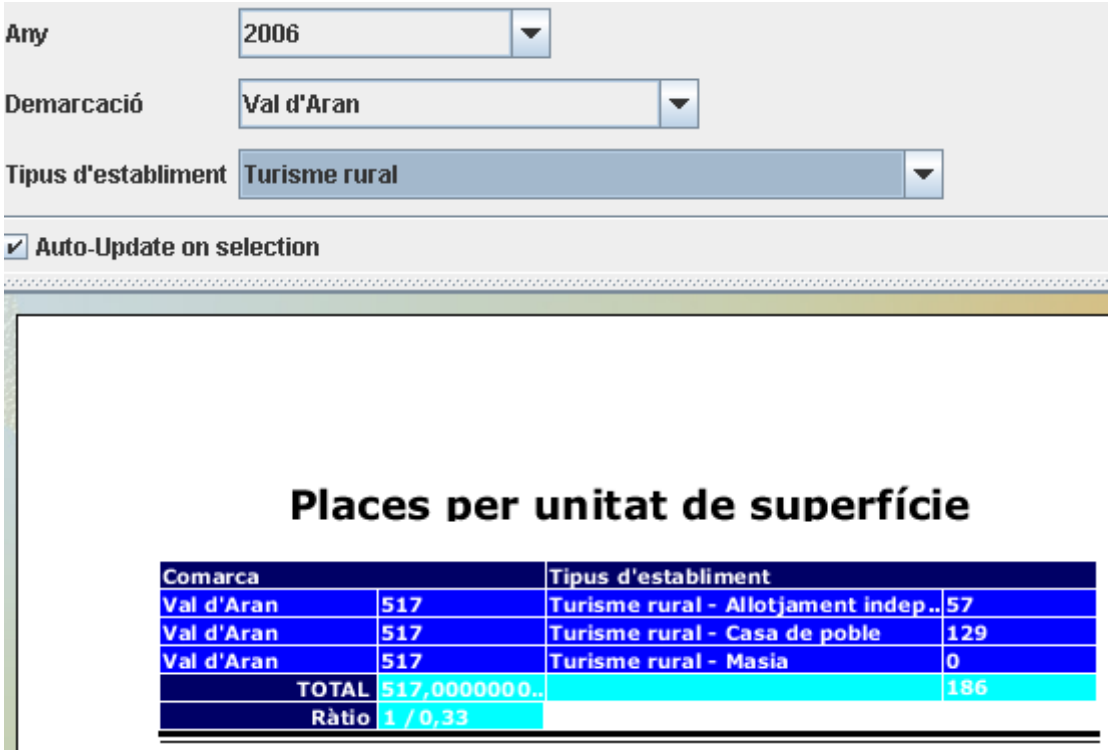
Es mostra una ràtio que representa el nombre d'habitants que hi ha per cada equipament públic d'un tipus determinat per un any i una demarcació concrets.



Il·lustració 44: Informe d'equipaments vs població

10. Quantitat de places ofertes / superfície del territori

Es mostra una ràtio que representa el nombre de places disponibles per un tipus d'establiment determinat per quilòmetre quadrat de superfície, segons l'any i la demarcació sol·licitats.


Any
2006
Demarcació
Val d'Aran
Tipus d'establiment
Turisme rural
Auto-Update on selection
Places per unitat de superfície
Comarca
Tipus d'establiment
Val d'Aran 517 Turisme rural - Allotjament indep.. 57
Val d'Aran 517 Turisme rural - Casa de poble 129
Val d'Aran 517 Turisme rural - Masia 0
TOTAL 517,0000000.. 186
Ràtio 1 / 0,33

5. Treball Futur

- Informes: Ús de *Dashboards*

Els *dashboards* són panells de control, on l'usuari té tota la informació mes important o enllaços per arribar a ella.

- Informes: Ús d'OLAP

Aquests informes permeten que l'usuari els manipuli per tal d'extreure nou coneixement potencialment útil, consultant les dades amb total llibertat i versatilitat. Es tracta en definitiva de dotar al sistema de més intel·ligència que permeti treure un avantatge competitiu quan la necessitat o les circumstàncies així ho indiquin.

- Informes: Incorporar gràfics als informes

Als informes estàtics generats amb PRD el gràfic s'ha d'incloure al propi disseny de l'informe i no és possible fer-ho des del portal. En el cas del informes dinàmics, tan sols es tractaria de prémer l'opció corresponent i ajustar els paràmetres desitjats, tasca que pot realitzar el propi usuari i dota de potencia afegida al sistema d'intel·ligència de negoci.

- Optimitzar aquells informes pels quals el client estigui més interessat.

En el cas dels informes que involucren una categoria d'equipaments com podria ser cultura o esports, i no un tipus d'equipament concret com podrien ser les biblioteques dintre de la categoria de cultura, es podrien fer passos per fer aquest tipus de consultes més ràpides a partir de camps precalculats que s'emmagatzemin a la base de dades, a costa d'un augment en el cost espacial del desenvolupament que seria assumible.

6. Conclusions

S'ha vist que avui en dia, per empreses de certa grandària, els magatzems de dades són imprescindibles. Els hi permeten obtenir nou coneixement a partir de les dades que inicialment tenen repartides entre diferents sistemes i diferents ubicacions.

Centrat en aquest treball de fi de grau, es pot considerar que l'etapa d'estudi i disseny del sistema sembla la més important pel correcte desenvolupament del projecte i el seu rendiment posterior de cara al client. En aquest cas, cal afegir el fet d'aprendre un nou paquet, en totes les fases del procés, des de la carrega amb el PDI, la creació del cub amb el Schema Workbench, com sobretot la generació d'informes amb el PRD, ha jugat un paper fonamental al resultat d'aquest treball, que no ha estat tot alló que l'autor d'aquest projecte voldria.

Tot i aquests contratemps, els objectius inicials que es plantejaven han estat assolits en tot el seu conjunt. Aquests objectius eren:

- 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.
- Perfeccionar la gestió i seguiment de projectes.
- Entregar un producte final on sigui mesurable aquest aprenentatge, i que compleixi els requeriments i funcionalitats demanats pel client.

Es fa notar que el domini de l'eina per crear els processos ETL és limitat, i que la falta de temps lliure per haver de compaginar quatre assignatures amb garanties no ha permés que s'aprofundís en el domini

d'aquesta buscant literatura relacionada amb la qual experimentar i millorar la capacitat d'ús. A més, aquesta falta d'ús ha jugat un paper important en el temps que s'ha trigat per confeccionar aquests processos ETL.

Queda un llarg camí per endavant per tal d'ampliar aquests coneixements, amb la relació de la intel·ligència de negoci amb la mineria de dades, i amb la Big Data.

7. Annexos

7.1 Annex A1 – Bibliografia

Enginyeria del Programari- XP03/05060/02078 - Universitat Oberta de Catalunya
Treball de Final de Carrera - Redacció de textos cientificotècnics, Presentació de documents i elaboració de presentacions - XP08/19018/00443 - Universitat Oberta de Catalunya

7.1.1 Enllaços d'interès a Internet

- http://dev.mysql.com
- http://forums.pentaho.com
- http://community.pentaho.com