TREBALL FINAL DE CARRERA

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

Memòria

Alumne: José Manuel Moreno Sánchez Consultor: Carles Llorach Rius

TFC Magatzem de dades
Enginyeria tècnica d'informàtica de Gestió
Universitat Oberta de Catalunya
Curs 2012-2013/2
17/06/2013

1 Resum

Aquest TFC de Magatzem de dades té com a tema a desenvolupar la “Construcció i explotació d’un magatzem de dades per a l’anàlisi d’informació sobre allotjaments turístics”. El projecte ens permetrà obtenir les informacions que l’Observatori Nacional d’Ocupació (ONdO) ens demana als requeriments del producte. El tipus d'ocupació que estudia ONdO és l'ocupació dels establiments turístics, per exemple el nombre de places ocupades als hotels. El que farà el producte del projecte serà l'anàlisi de l'evolució d'allotjaments turístics a Catalunya i les possibles correlacions entre allotjaments i equipaments públics. Per exemple nombre d'establiments i nombre d'equipaments en una zona. Cal fer notar que a partir de les relacions de les dades obtindrem informació. ONdO ens proporciona unes dades que han de ser transformades per poder ser aprofitables per a l'ús demanat, ja que estan en diferents formats i expressades de formes diverses, per exemple les dades de població en unitats, milers i milions. Per a aconseguir-ho he usat eines ETL (*extract transform and load*), que permeten usar fonts de dades diferents, transformar-les i carregar-les en una base de dades. La base de dades és MySQL, a la qual he fet procediments emmagatzemats que permeten obtenir les dades requerides. Una vegada fet això les dades es mostren usant informes amb paràmetres. Les eines amb que ho he fet són del conjunt d'eines de Pentaho.

2 Índex de continguts

3 Cos de la memòria

El cos de la memòria descriu el treball fet al llarg del TFC.
A continuació tenim una breu introducció sobre diferents aspectes del TFC.

3.1 Introducció

Aquest capítol de la memòria pel Treball de Fi de Carrera (TFC) integra el treball fet a les PAC's al llarg del semestre 2012-2013-2 a l'àrea de Magatzem de Dades.
Els subapartats d'aquesta introducció expliquen de forma resumida el que s'ha treballat al llarg del semestre.

3.1.1 Justificació del TFC i context en el qual es desenvolupa: punt de partida i aportació del TFC

El projecte està justificat per la necessitat de fer possible l'obtenció d'una informació basada en dades disperses com a punt de partida. Una vegada aplicades les tècniques de magatzems de dades en aquest TFC es podrà treure profit d'elles amb aquesta informació resultant.

Aquestes tècniques d'aprofitament de dades en diferents formats, a vegades provinents d'una base de dades i a vegades en altres formats, resulten molt interessants per a l'entorn de l'empresa, ja que estant integrades, aquestes dades poden donar lloc a informació que li doni a l'empresa avantatges competitius.

3.1.2 Objectius del TFC

Obtenir el producte demanat per ONdO, per a fer-ho es segueix el descrit en aquesta memòria sobre el treball realitzat al llarg del semestre.
Posar en pràctica els coneixements relacionats amb les Bases de dades adquirits durant els estudis.
Obtenir experiència en el desenvolupament de projectes com a resultat del proces d'elaboració del TFC, que es fa partint dels requeriments del client. I obtenir experiència també en la planificació de les tasques a realitzar.

3.1.3 Enfocament i mètode seguit

Aquest projecte s'ha fet amb la metodologia de cicle de vida iteratiu i incremental basat en el cicle de vida en cascada. Mes endavant s'explica a la introducció de l'apartat d'anàlisi i disseny.

3.1.4 Planificació del projecte

Un projecte necessita complir uns objectius en un temps determinat. En aquest apartat es descriurà el problema que es pretén resoldre, les eines que s'usen, el treball concret que es porta a terme i la seva descomposició en tasques i fites temporals.

3.1.5 Productes obtinguts

El producte obtingut és el conjunt d'informes encarregats per ONdO. S'explica a l'apartat sobre

implementació. Breument es pot dir que els informes són visibles per als usuaris que tenen permisos per veure'ls, tenen paràmetres que permeten mostrar les dades que ens interessin i estan integrats a Pentaho BI Platform, que permet la seva visualització en format web.

3.1.6 Breu descripció dels altres capítols de la memòria

A continuació tenim el que s'ha treballat al llarg del semestre en la planificació, l'anàlisi i disseny, i la implementació. Aquests apartats són una adaptació a la memòria de les entregues de les PAC's.

3.2 Planificació

3.2.1 Introducció

Aquesta part del document descriu el pla de treball per l'elaboració del treball final de carrera en l'àrea de magatzems de dades. Té com a tema la construcció i explotació d'un magatzem de dades per a l'anàlisi d'informació sobre allotjaments turístics.

Es descriu el problema que es pretén resoldre, les eines que s'usaran, el treball concret que es portarà a terme i la seva descomposició en tasques i fites temporals.

El *Data Warehouse* (magatzem de dades), ens permetrà integrar diferents fonts de dades, que poden ser molt diverses. Posteriorment aquestes dades es podran analitzar i es podrà extreure informació útil per a la presa de decisions.

3.2.2 Justificació del projecte (idoneïtat)

El projecte està justificat per la necessitat de fer possible l'obtenció d'una informació basada en dades disperses. Una vegada aplicades les tècniques de magatzems de dades es podrà treure profit d'elles amb aquesta informació resultant.

Aquestes tècniques d'aprofitament de dades en diferents format, a vegades provinents d'una base de dades i a vegades en altres formats, resulten molt interessants per a l'entorn de l'empresa, ja que estant integrades aquestes dades poden donar lloc a informació que li doni a l'empresa avantatges competitius.

3.2.2.1 Per què el projecte?

En aquesta àrea de TFC es pretén que l'estudiant s'introdueixi en aquest camp dins del marc de les bases de dades que representa un repte en la gestió de la informació.
Les empreses tenen un gran volum de dades emmagatzemades en els seus sistemes operacionals unes vegades, i altres la rellevant per a la presa de decisions. El magatzem de dades és un repositori d'informació orientat a recopilar, resumir i tractar eficientment aquestes dades de forma que faciliti l'anàlisi d'informació des de diverses perspectives o dimensions d'anàlisi, una d'elles és la perspectiva de màrqueting.

La forma de fer-ho és la següent: el proces ETL (*Extract, transform and load*) farà l'extracció de les dades del sistema de l'empresa, la transformació i la càrrega al magatzem de dades (DW). El magatzem de dades i les eines OLAP permetran l'explotació en temps real (*on-line*) de les dades, l'anàlisi de les quals portarà a conclusions útils per a la presa de decisions, ja que reflectirà tendències i això podrà ser aprofitat per tenir avantatge sobre els competidors. Aquí és a on entra en joc el *Business Intelligence*.
Tenim diferents formes d'obtenir la informació depenent del moment en que aquesta es produeixi i la tinguem disponible:

- que ha passat (*reporting* operatiu i anàlisi OLAP)
- que està passant (*dashboards*)
- que vull que passi (quadres de comandament estratègics i/o operacionals)
- que passarà (*data mining*).

3.2.2.2 Descripció del projecte

El projecte ens permet obtenir les informacions que l'Observatori Nacional d'Ocupació (ONdO) ens demana als requeriments del producte. El tipus d'ocupació que estudia ONdO és l'ocupació dels establiments turístics, per exemple el nombre de places ocupades als hotels. El que fa el producte del projecte serà l'anàlisi de l'evolució d'allotjaments turístics a Catalunya i les possibles correlacions entre allotjaments i equipaments públics. A partir de les relacions de les dades obtenim informacions útils per a la presa de decisions.

Per tal de tenir un desenvolupament ordenat i ben documentat, dividim el projecte en quatre etapes:

PAC1 --> **Pla de Treball**
On s'ha d'indicar amb un cert nivell de detall les tasques que s'hauran de realitzar, juntament amb una anàlisi de riscos i un diagrama de Gantt.

PAC2 --> **Anàlisi i Disseny**
S'ha de fer l'anàlisi de la problemàtica proposada a l'enunciat i s'ha de fer el disseny de la solució generant el model multidimensional de les dades, així com tota la informació necessària per a poder assolir els objectius demanats.

PAC3 --> **Implementació**
S'ha de crear la BDD tal i com s'ha dissenyat a l'apartat anterior i també s'haurà de carregar les dades de les que es disposa. En aquest procés de càrrega hi ha d'haver un procés de transformació i adequació al que es demana. Un cop carregada la BDD s'han de generar els informes necessaris per a poder resoldre el que s'ha demanat. Aquests informes s'han de generar amb l'eina indicada pel consultor.

Lliurament final (inclou memòria, producte i presentació virtual)
Cal crear la memòria del TFC on s'explica el treball realitzat (habitualment serà una suma i adequació dels entregables anteriors). També s'ha de crear una presentació virtual on s'expliqui el treball fet. Aquesta presentació serà la que permetrà la defensa del treball davant del tribunal.

3.2.3 Objectius del projecte

L'objectiu del projecte és la posada en pràctica d'uns coneixements adquirits durant els estudis, en aquest cas en l'àrea de bases de dades. I a més en el meu cas, com que no tinc experiència professional en el tema de magatzems de dades és un treball de recerca de coneixements sobre els magatzems de dades i l'ús del programari relacionat. En el proces s'ha d'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 i altres fonts de dades.

3.2.3.1 Generals

Els objectius generals del projecte son:
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 i altres fonts.
L'anàlisi de tècniques existents per a projectar la base de dades d'un DW. El disseny estarà orientat a un magatzem de dades físic ROLAP, creant les dimensions, atributs i fets necessaris. Han de considerar-se els factors: desnormalització de taules, inclusió d'informació agregada, historificació de la informació, etc.
Aprofundir en aspectes concrets com les tècniques de tractament de dades i la seva integració en el model de dades físic del DW.
Obtenir experiència en el desenvolupament de projectes com a resultat del proces d'elaboració del TFC, que es fa partint dels requeriments del client i les dades aportades.
Planificació de les tasques a realitzar.
Obtenir experiència en l'elaboració d'informes.
Obtenir un producte, una memòria i una presentació virtual.

3.2.3.2 Específics

Aplicar un proces ETL als fitxers de dades aportats per ONdO per crear un *Data Warehouse* que ens permeti obtenir i visualitzar aquesta informació dintre d'una temporalitat a nivell d'any:

- 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

Tota aquesta informació es podrà consultar de forma agregada, per comarca/província, tipus d'establiment i categoria.
Es crearà un conjunt d'informes a on es mostrin aquestes informacions.
Al proces ETL s'ha de tenir en compte que les fonts de dades provenen de sistemes diferents i estan en formats diversos, s'han de tractar per poder usar-les correctament. A més s'ha de calcular el

nombre d'habitants com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent. S'haurà de fer arrodoniment per obtenir nombres sencers.

S'ha de produir la documentació de les 3 PACs, la memòria que expliqui el proces de creació del projecte i una presentació que sintetitzi el més rellevant.

3.2.4 Requeriments de la solució

Els requeriments estan descrits a l'enunciat, i mostren el que espera el client del producte final, que en aquest cas és l'Observatori Nacional d'Ocupació (ONdO). El que desitja és aprofundir en l'evolució dels establiments turístics, esmenta el creixement del seu nombre i als arxius aportats ens dona xifres. També desitja analitzar les possibles correlacions entre allotjaments i equipaments públics. Per a fer això es construeix i s'explota un magatzem de dades.
En tractar-se d'un TFC tenim alguns requeriments més. Hem de fer una memòria que expliqui el treball que s'ha fet, i una presentació que mostri el més destacable del proces de desenvolupament del TFC i els seus resultats.
L'accés a la informació requerida pel client es fa mitjançant Pentaho BI Platform.

3.2.4.1 Funcionals

Les dades que el client ens aporta mitjançant uns fitxers les hem de passar per un proces d'extracció, transformació i càrrega. A aquest proces se l'anomena ETL (*Extract, Transform and Load*). El resultat d'aquest proces és una base de dades relacional preparada per ser utilitzada per tractar la informació que mostraran els informes.

Usuaris
Per tal de mantenir una seguretat en l'accés a la informació i la plataforma Pentaho BI hi hauran 3 tipus d'usuari. 2 Usuaris finals; un que recopila informació bàsica (principalment informes) per després distribuir-los entre els membres d'ONdO, i un altre usuari avançat que podrà dissenyar noves consultes i explorar les dades més en profunditat segons les necessitats del moment per treure tot el profit possible. El tercer usuari és de tipus tècnic i s'encarrega de revisar les càrregues, resoldre les incidències, gestionar els usuaris i els permisos.

3.2.4.2 No funcionals

El sistema té aquestes propietats que el defineixen.

Seguretat
Pentaho BI Platform disposa d'un sistema de seguretat en l'accés a la informació que serà utilitzat pel portal per permetre aquest accés als usuaris autoritzats a cada un dels informes.

Usuaris	**Funcions**
Bàsic	recopila informació bàsica, principalment informes.
Avançat	pot analitzar les dades lliurement per extreure informació el més valuosa possible, especialment la que es troba analitzant les dades.

Administrador	fa les càrregues, resol les incidències, gestiona els usuaris i els permisos d'accés a la informació que tenen els usuaris.

Portabilitat
En tractar-se d'un sistema que presenta la informació en format web, s'ha de disposar d'un navegador web per veure la informació.

Rendiment
Un Magatzem de dades té com a característica que les dades que es consulten s'han de mostrar de forma immediata. Si s'han de mostrar dades calculades han d'estar prèviament calculades o ser càlculs prou ràpids.

Facilitat d'ús
L'usuari bàsic no ha de tenir cap coneixement especial per recopilar informació bàsica, que principalment seran informes. L'usuari avançat ha de tenir coneixements especialitzats perquè ha de poder analitzar les dades lliurement per extreure informació el més valuosa possible, especialment la que es troba analitzant les dades.

Fiabilitat
El sistema està construït usant MySQL i Pentaho. La fiabilitat és la proporcionada per ells i és prou bona com per ser molt usats.
El Centre de Software de Ubuntu diu això sobre MySQL.

> MySQL is a fast, stable and true multi-user, multi-threaded SQL database server. SQL (Structured Query Language) is the most popular database query language in the world. The main goals of MySQL are speed, robustness and ease of use.

A http://www.pentaho.com/partners/technology/ podem veure això:

> Pentaho Technology Partners work with us to ensure sustainable, supportable integration of our products, making a wide range of compatible technologies available to customers. Pentaho's established relationships with these partners enable us to work together to support integration points and resolve joint customer issues.

3.2.4.3 Anàlisi de les dades

ONdO ens proporciona les dades en arxius de tipus csv i txt que s'han de revisar i normalitzar perquè tenen els problemes que a continuació es detallen.

Els csv tenen diferents tipus de separador i codificació.
A l'arxiu poblacio.csv les dades dels anys 2012 i 2006 estan en milers, excepte per a Barcelona al 2006 que està en milions, i per al següents municipis petits al 2006 les dades estan en unitats: Botarell, Cabra del Camp, Cabrera d'Anoia, Castelldans, Castellfollit de la Roca, Castellnou de Bages, Cornudella de Montsant, Flaçà, Gualba, Guardiola de Berguedà, Les, Montferrer i Castellbò, Navata, Perafort, Pla del Penedès (El), Port de la Selva (El), Puigpelat, Riudecanyes, Sant Llorenç de Morunys, Sant Martí de Centelles, Santa Eulàlia de Riuprimer, Serinyà, Ullà i Viladrau.

Les dades de població per al 2012 per sexe tenen el mateix problema de milers i unitats barrejades per a diferents poblacions.
Per a Canonja (La) no hi ha dades del 2007 al 2011 i per al 2006 hi ha població 0.

Els txt, referits als establiments turístics entre 2006 i 2012 tenen altres tipus de diferències entre ells que s'hauran de tractar per normalitzar-les:

- A l'any 2012 no tenim constància de que les dades siguin a data 31 de desembre, com ho són dels altres anys.
- Advertiment: L'11 de maig del 2010 s'han revisat les sèries 2007-2008.
- A partir de l'any 2011, a l'article 49 del Decret 183/2010, de 23 de novembre, d'establiments d'allotjament turístic, la capacitat d'allotjament d'un càmping en nombre de places s'obté multiplicant per tres el nombre total d'unitats d'acampada. Aquest canvi comporta un trencament de la sèrie que n'impedeix la comparació interanual, ja que per als anys anteriors el nombre de places d'un càmping s'obtenia multiplicant per dos i mig el nombre total d'unitats d'acampada.
- Hi ha diferències entre els arxius en el nombre de columnes buides.
- A l'any 2012 les columnes de les estrelles no hi son, hi ha sobre hotel i sobre hostals o pensions.
- A l'any 2006 tenim establiment privats però no de luxe, a la resta d'anys tenim establiments de luxe però no privats.
- A l'any 2006 tenim les columnes: allotjament independent, masia i casa de poble. A la resta dels anys tenim: casa de poble compartida, casa de poble independent, masia i masoveria.
- A l'any 2012 les dades de categoria estan 2 files cap amunt.
- A l'any 2012 tenim dades de Turisme Rural a partir de les dades del Departament d'Empresa i Ocupació.

Com a requisit no funcional relacionat amb el TFC tenim que l'alumne ha d'adquirir uns coneixements teòrics i de maneig de programes que li permetin desenvolupar el treball satisfactòriament. A més ha de fer les entregues assenyalades a l'apartat 3.2.2.2 Descripció del projecte.

Qualitat de dades

Correcció de dades errònies
He corregit les dades de població per tenir-les totes en unitats, amb l'eina Pentaho Data Integration, amb una funció que permet fer-ho en llenguatge Java Script.
Les dades de geolocalització he pres la decisió de no tocar-les perquè no tenia dades millors i deixant-les com estan es fa evident que estan malament.

Incorporació de dades que no venien en els arxius inicials
No he incorporat dades, però si he descrit la forma de fer-ho i ho deixo com treball futur. Sí que he inclòs identificadors d'àrea i comarca a les dades d'establiments i equipaments.

Dades descartades per incompletes
Hi ha dades incompletes sobre la població "La Canonja", he descrit en aquesta memòria com tractar-les a l'apartat 3.3.7

Volum de dades

La taula que conté més files inicialment és la d'equipaments, 31771.
La taula de municipis una vegada normalitzada conté 32076 files, inicialment tenia 468.
Les taula d'establiments per al 2006 una vegada normalitzada conté 1696 files, inicialment tenia 63.

3.2.5 Funcionalitats a desenvolupar

Obtenció del magatzem de dades mitjançant processos ETL.
Desenvolupament dels següents informes:

- 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

3.2.6 Resultats esperats

S'espera obtenir un magatzem de dades i uns informes que compleixin els requeriments del client.
S'espera complir la planificació per no anar amb retards. Per tal de fer més controlable la correcta finalització a temps del projecte, les dates de les parts a entregar (apartat 3.2.8.3) es tindran molt en compte. Si hi hagués algun retard es tindrà en compte l'anàlisi de riscos per complir les dates.

3.2.7 Organització del projecte

En l'organització del projecte és essencial que els components *software* i *hardware* siguin capaços de complir amb les tasques que s'esperen d'ells. A continuació els detallem.

3.2.7.1 Components SW / HW

Components SW (*software*)
Habitualment faig servir sistemes operatius Ubuntu amb *software* gratuït que aquest ofereix al seu centre de *software*.
Usaré de LibreOffice: el processador de textos per fer les PACs i la memòria, LibreOffice Calc per obrir i editar arxius de l'Excel, i LibreOffice Impress per fer la presentació.
L'editor de diagrames Dia.
El gestor de projectes Planner per a fer el diagrama Gantt.
RecordMydesktop per fer vídeos, útil per fer preguntes sobre dubtes i per fer la presentació final.
Avidemux per editar vídeo.
Mezclador ALSA de Gnome per controlar les entrades de so de les gravacions de vídeo.
Transcodificador de vídeo Transmageddon per passar del format de vídeo .ogv a .mp4

KompoZer per crear i editar pàgines web, útil per la seva característica WYSIWYG (*What You See Is What You Get*).
Xarchiver i Gestor de archivadores per comprimir i descomprimir fitxers.
Defraggler Portable per desfragmentar la màquina virtual.
FileZilla com a client FTP per pujar els arxius de la màquina virtual en l'entrega.
Mozilla Firefox per veure la plataforma BI, els informes i sortides de dades i la consola d'administració.
VirtualBox per la màquina virtual que conté el sistema operatiu sobre el que es treballa.

L'àrea del TFC proporciona una màquina virtual de virtualBox amb tot el programari necessari, tot open-source per evitar que tingui cost i així facilitar el seu accés a empreses de grandària mitjana.

La màquina virtual conté els següents programes:
· Servidor web Tomcat per veure les pàgines web dels informes i els programes en format web.
· MySQL com a base de dades.
· MySQL Workbench per crear i gestionar la base de dades.
· Keetle per fer els processos ETL.
· Servidor BI (Pentaho BI Server de la suite Pentaho).
· *Dashboard* (CDE) per fer quadres de comandament operatiu.
· SAIKU per fer anàlisi OLAP.

A la suite Pentaho tenim:
· Pentaho Data Integration. Per fer una visualització inicial de les taules i els processos ETL (*extract, transform and load*).
· Report Designer per crear informes.
· Administration Console (localhost:8099) per crear i gestionar usuaris i rols per a la visualització d'informes.
· Pentaho BI Platform (localhost:8080) per crear i visualitzar informes.

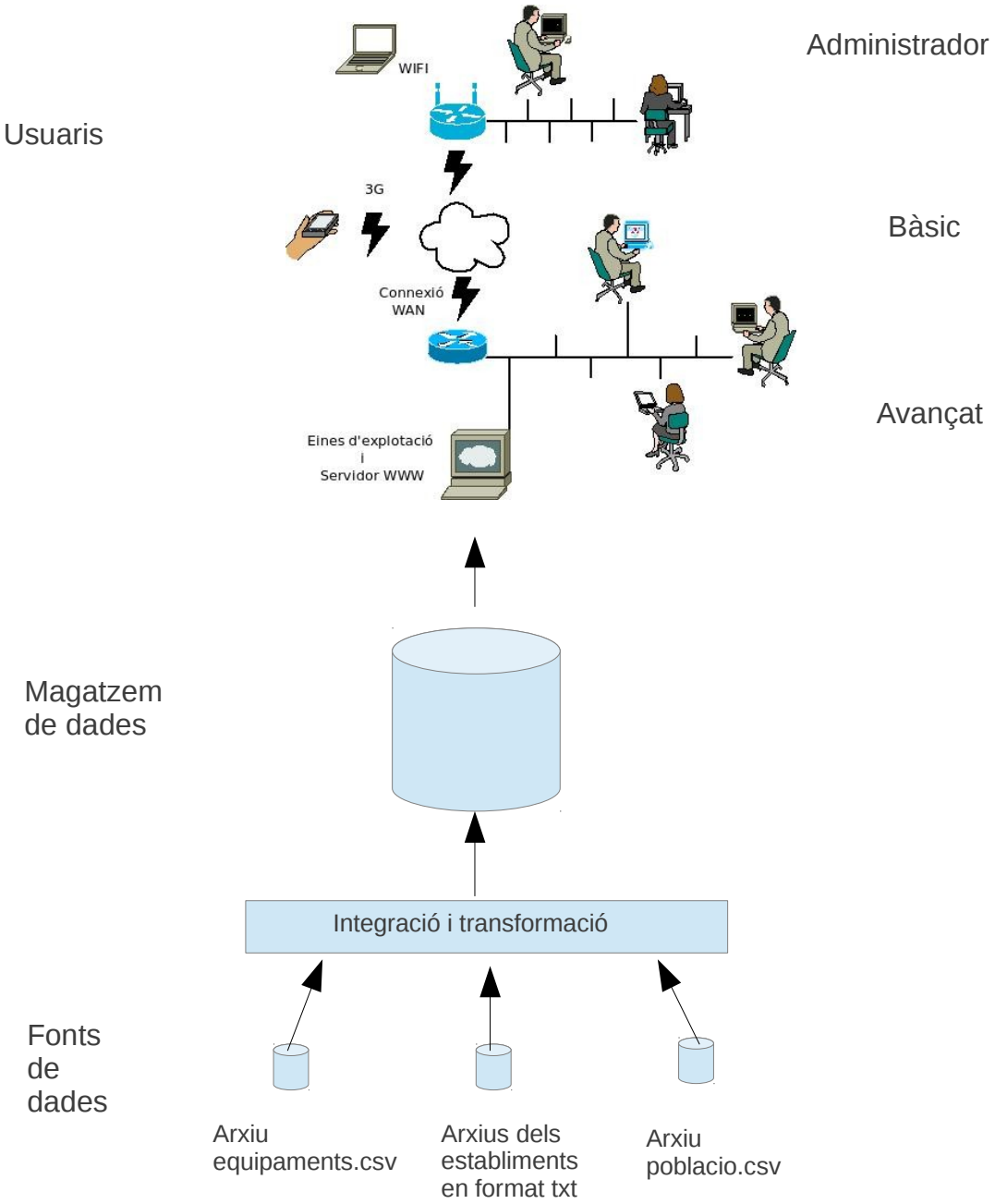
· Schema Workbench per generar els arxius MDX del cub al que es connectarà OLAP per fer un anàlisi. Cal pujar l'arxiu MDX al Pentaho BI Server (localhost:8080), Pentaho BI Server té un assistent OLAP que també es pot usar per fer això.

Components HW (*hardware*)
Com que es farà servir una màquina virtual caldrà que no anem massa justos de recursos.
La màquina virtual pot ocupar uns 9 GB del disc i necessita memòria RAM, cal tenir-ho en compte.
Per obtenir bons resultats en la velocitat de la màquina virtual se li pot assignar 3.9 GB de RAM i usar un disc dur sòlid.
Una memòria USB pot emmagatzemar les còpies de seguretat.

3.2.7.2 Arquitectura del projecte

Les dades operacionals que ONdO ens subministra es carreguen via ETL al magatzem de dades, que és el "*Sistema informàtic utilitzat com a eina de suport per a la presa de decisions que integra una gran diversitat d'informació procedent de diferents bases de dades i que permet realitzar consultes complexes i de tipus analític sobre aquesta informació*". Utilitzant *Bussiness Intelligence*, que és "*el conjunt d'estratègies i eines enfocades a l'administració i creació de coneixement*

mitjançant l'anàlisi de dades existents en una organització o empresa", els usuaris poden accedir a la informació fent consultes complexes i de tipus analític sobre la informació sense adonar-se de la complexitat.
El magatzem de dades no conté dades volàtils com tindria una base de dades operacional, no podem formar una "pel·lícula" de les dades al llarg del temps, la temporalitat que tindrem serà anual. Al producte tindrem una càrrega única amb tot l'històric.



L'usuari accedeix a la informació a través de Pentaho BI Platform.
L'usuari accedeix a la informació a través d'una interfície web que funciona des de el mateix ordinador a on està el Data Warehouse, la xarxa local o inclús es podria intentar fer la connexió des

de Internet fent la configuració del router, IP tables, xarxa de la màquina virtual, les propietats de la connexió i del firewall. El funcionament d'accés des de Internet no s'ha arribat a comprovar.

3.2.7.3 Tecnologies a utilitzar

Interfície web.
Base de dades MySQL.
La suite Pentaho ens oferirà programari per a les següents tasques:

- ETL per carregar el Data Warehouse.
- *Bussiness Intelligence* per fer consultes complexes i de tipus analític sobre la informació.
- OLAP per al processament analític de les dades.

3.2.7.4 Anàlisi de riscos

Al llarg del projecte poden sorgir incidències que facin perillar el normal desenvolupament o l'acompliment de la previsió.

Pla de contingències	
Risc	**Acció correctiva**
Haver de dedicar temps del projecte a altres tasques	s'intentaria compensar aquesta falta de dedicació el mes aviat possible.
Avaria a l'ordinador habitual	s'intentarà aconseguir una altra màquina el mes aviat possible.
Deixa de funcionar el sistema operatiu habitual	es tindrà preparat un altre amb els programes necessaris per continuar.
Pèrdua d'informació	es faran copies de seguretat de la feina feta, si son arxius de text inclús varies vegades al dia. Aquesta copia es farà sobre una memòria USB per fer-la més àgil.
Per motius de salut no s'acompleix amb el previst	s'intentarà compensar el més aviat possible.
Hi han retards puntuals	s'intentarà compensar sacrificant alguna hora de son.
L'aprenentatge dels programes crea problemes en l'avenç	es tindran alternatives en què treballar per evitar consumir el temps sense aconseguir fites.

3.2.8 Proposta d’activitats i cronograma

Al llarg del TFC tenim una trobada virtual, 3 entregues de PACs que ens permetran tenir un correcte seguiment i una entrega final composta per memòria, producte i presentació. La trobada amb data a consensuar i les entregues seguint un calendari.

3.2.8.1 Relació d’activitats

Al llarg del projecte es fan unes tasques que concreten el consum del temps total disponible.

Nº	**Fita**	**Tasca**	**Durada**	**Data inici**	**Data final**	**Precedències**
1	Preparació inicial					
2		Llegir el Pla Docent	1 h	27/02/2013	27/02/2013	
3		Descarregar els materials de l'assignatura	2 h	27/02/2013	27/02/2013	2
4		Llegir el mòdul Redacció de textos cientificotècnics	6 h	28/02/2013	01/03/2013	3
5		Llegir el mòdul Presentació de documents i elaboració de presentacions	6 h	02/03/2013	03/03/2013	4
6	Trobada virtual					
7		Elegir data	5 min	06/03/2013	06/03/2013	
8		Crear compte gmail	10 min	06/03/2013	06/03/2013	
9		Instal·lar el navegador Chromium i el plugin per Hangouts	3 h	09/03/2013	09/03/2013	
10		Assistència a la reunió	1 h30 min	10/03/2013	10/03/2013	
11	PAC1	**Pla de Treball**		28/02/2013	12/03/2013	3
12		Llegir l'enunciat del projecte	1 h	28/02/2013	28/02/2013	
13		Mirar els fitxers de dades	2 h	28/02/2013	28/02/2013	12
14		Lectura de materials complementaris	2 h	09/03/2013	09/03/2013	
15		Llegir PACs anteriors d'exemple	2 h	04/03/2013	04/03/2013	5
16		Elaboració de la PAC1	8 d 1h	04/03/2013	12/03/2013	15
17		· Justificació, objectius i requeriments	3 d	04/03/2013	07/03/2013	
18		· Organització del projecte	3 d	07/03/2013	10/03/2013	
19		· Activitats i cronograma	2 d 1h	10/03/2013	12/03/2013	
20	PAC2	**Anàlisi i Disseny**		13/03/2013	16/04/2013	11
21		Descarregar la màquina virtual i fer la instal·lació	2 h	13/03/2013	13/03/2013	
22		Tornar a llegir l'enunciat	30 min	13/03/2013	13/03/2013	
23		Lectura de materials complementaris	20 h	13/03/2013	19/03/2013	22
24		Lectura de PACs anteriors d'exemple	2 h	19/03/2013	20/03/2013	23
25		Elaboració de la PAC	27 d	20/03/2013	16/04/2013	24
26		· Disseny BD	15 d	20/03/2013	04/04/2013	
27		· Anàlisi processos ETL	12 d	04/04/2013	16/04/2013	
28	PAC3	**Implementació**		17/04/2013	29/05/2013	20
29		Tornar a llegir l'enunciat	30 min	17/04/2013	17/04/2013	
30		Lectura de materials complementaris	20 h	17/04/2013	23/04/2013	29
31		Aprenentatge del programari	20 h	23/04/2013	30/04/2013	30
32		Lectura de PACs anteriors d'exemple	4 h	30/04/2013	01/05/2013	31
33		Elaboració de la PAC	28 d	01/05/2013	29/05/2013	32
34		· Creació BD	4 d	01/05/2013	05/05/2013	
35		· Processos ETL	10 d	05/05/2013	15/05/2013	

36		· Creació d'informes	14 d	15/05/2013	29/05/2013	
37	Lliurament final			30/05/2013	17/06/2013	28
38		Memòria	10 d	30/05/2013	08/06/2013	
39		Producte	4 d	09/06/2013	12/06/2013	38
40		Presentació	5 d	13/06/2013	17/06/2013	39
41	Defensa		4 d	18/06/2013	21/06/2013	37

Això és una planificació que serveix per veure si anem bé de temps i poder preveure que s'han de fer esforços extra per arribar als objectius.

3.2.8.2 Estimació de temps

Es treballarà tres hores al dia, tots els dies. Si per qualsevol causa no s'acompleix l'assoliment de la càrrega de feina s'allargaria la jornada.

3.2.8.3 Fites a complir

S'han de complir les fites marcades al pla d'estudis de l'assignatura. Son les següents:

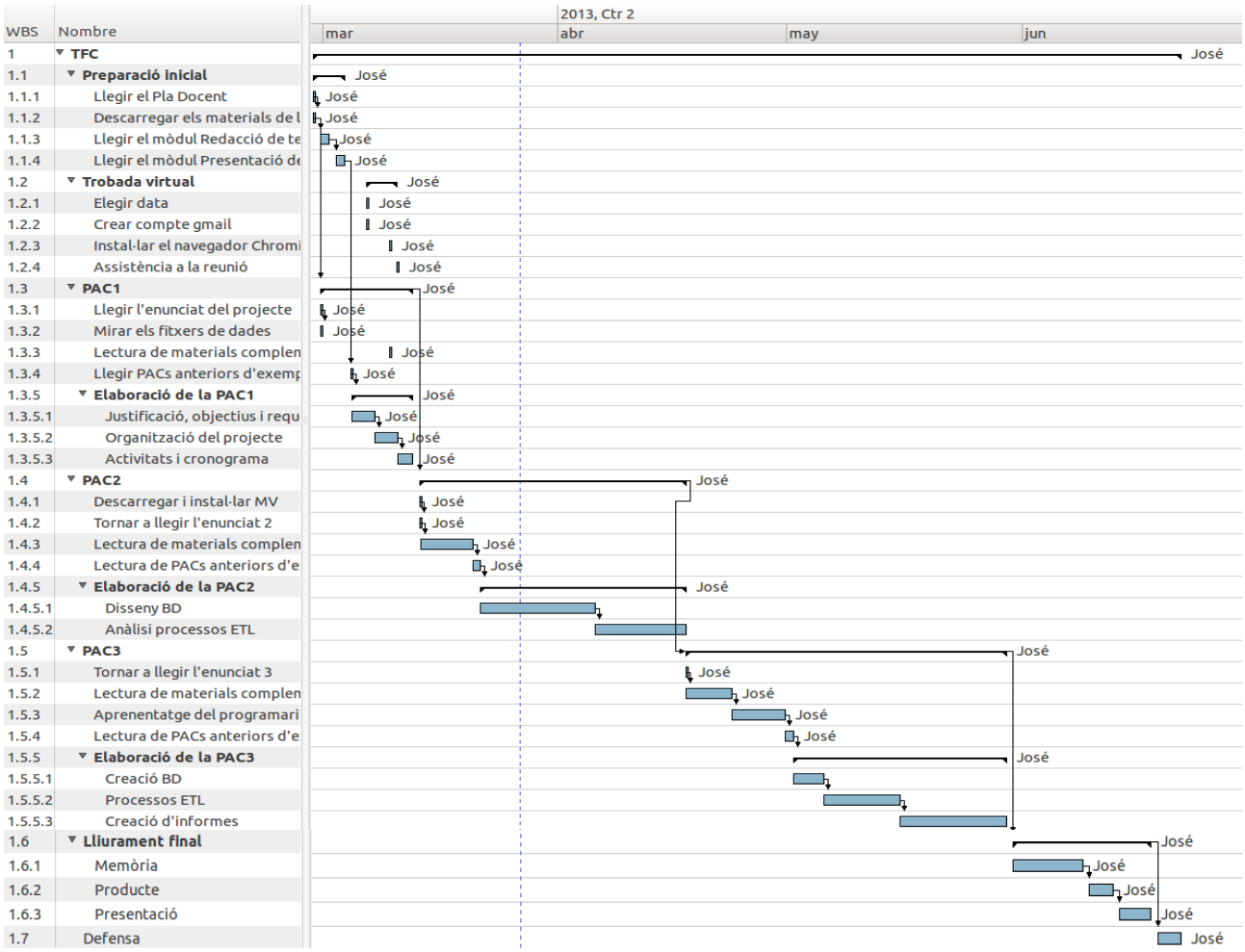
Títol	**Inici**	**Lliurament**
PAC1	28/02/2013	12/03/2013
PAC2	13/03/2013	16/04/2013
PAC3	17/04/2013	29/05/2013
Lliurament final i defensa	30/05/2013	17/06/2013

3.2.8.4 Diagrama de Gantt

Planificació de tasques

WBS	Nombre	Inicio	Fin	Duración
1	**TFC**	**feb 27**	**jun 21**	**115d**
1.1	**Preparació inicial**	**feb 27**	**mar 3**	**5d**
1.1.1	Llegir el Pla Docent	feb 27	feb 27	1h
1.1.2	Descarregar els materials de l'assigna	feb 27	feb 27	2h
1.1.3	Llegir el mòdul Redacció de textos cie	feb 28	mar 1	2d
1.1.4	Llegir el mòdul Presentació de docum	mar 2	mar 3	2d
1.2	**Trobada virtual**	**mar 6**	**mar 10**	**4d 1h**
1.2.1	Elegir data	mar 6	mar 6	5min
1.2.2	Crear compte gmail	mar 6	mar 6	10min
1.2.3	Instal·lar el navegador Chromium i el p	mar 9	mar 9	1d
1.2.4	Assistència a la reunió	mar 10	mar 10	1h 30min
1.3	**PAC1**	**feb 28**	**mar 12**	**13d**
1.3.1	Llegir l'enunciat del projecte	feb 28	feb 28	1h
1.3.2	Mirar els fitxers de dades	feb 28	feb 28	2h
1.3.3	Lectura de materials complementaris	mar 9	mar 9	2h
1.3.4	Llegir PACs anteriors d'exemple	mar 4	mar 4	2h
1.3.5	**Elaboració de la PAC1**	**mar 4**	**mar 12**	**8d 1h**
1.3.5.1	Justificació, objectius i requeriment	mar 4	mar 7	3d
1.3.5.2	Organització del projecte	mar 7	mar 10	3d
1.3.5.3	Activitats i cronograma	mar 10	mar 12	2d 1h
1.4	**PAC2**	**mar 13**	**abr 16**	**34d 1h**
1.4.1	Descarregar la màquina virtual i fer la instal·lació	mar 13	mar 13	2h
1.4.2	Tornar a llegir l'enunciat 2	mar 13	mar 13	30min
1.4.3	Lectura de materials complementaris	mar 13	mar 19	6d 2h
1.4.4	Lectura de PACs anteriors d'exemple	mar 19	mar 20	2h
1.4.5	**Elaboració de la PAC2**	**mar 20**	**abr 16**	**27d**
1.4.5.1	Disseny BD	mar 20	abr 4	15d
1.4.5.2	Anàlisi processos ETL	abr 4	abr 16	12d
1.5	**PAC3**	**abr 17**	**may 29**	**42d 2h**
1.5.1	Tornar a llegir l'enunciat 3	abr 17	abr 17	30min
1.5.2	Lectura de materials complementaris	abr 17	abr 23	6d 2h
1.5.3	Aprenentatge del programari	abr 23	abr 30	6d 2h
1.5.4	Lectura de PACs anteriors d'exemple	abr 30	may 1	1d 1h
1.5.5	**Elaboració de la PAC3**	**may 1**	**may 29**	**28d**
1.5.5.1	Creació BD	may 1	may 5	4d
1.5.5.2	Processos ETL	may 5	may 15	10d
1.5.5.3	Creació d'informes	may 15	may 29	14d
1.6	**Lliurament final**	**may 30**	**jun 17**	**19d**
1.6.1	Memòria	may 30	jun 8	10d
1.6.2	Producte	jun 9	jun 12	4d
1.6.3	Presentació	jun 13	jun 17	5d
1.7	Defensa	jun 18	jun 21	4d

Diagrama de Gantt



3.3 Anàlisi i disseny

3.3.1 Introducció

En aquest TFC de magatzem de dades en el que es treballa amb un tema desconegut del qual no hem rebut formació al llarg de la carrera i en el que desconeixem el software amb el qual hem de treballar, l'opció del mètode de cicle de vida iteratiu i incremental basat en el cicle de vida en cascada tindria justificació donat que probablement ens anirem trobant amb dificultats que forçaran a tirar enrere per fer correccions.

Al cicle de vida iteratiu i incremental es treballa sobre una part dels requisits passant per les fases del cicle de vida en cascada, però d'una forma més flexible, modificant el que calgui de parts anteriors. I després es fa el mateix amb una altra part. El que aconseguim és una part de la feina que sabem que està ben feta i ens serveix de base per a continuar amb una altra part.

Donat que el TFC està plantejat d'una altra manera, la falta de coneixement en profunditat de la temàtica i del programari faran difícil no cometre errors, probablement serà necessari tornar enrere per fer correccions sobre el que en aquesta PAC es planteja.

3.3.2 Requeriments funcionals / no funcionals

Els requeriments estan descrits a l'enunciat, i mostren el que espera el client del producte final, que en aquest cas és l'Observatori Nacional d'Ocupació (ONdO). El que desitja és aprofundir en l'evolució dels establiments turístics, esmenta el creixement del seu nombre i als arxius aportats ens dona xifres. També desitja analitzar les possibles correlacions entre allotjaments i equipaments públics. Per a fer això construirem i explotarem un magatzem de dades.
En tractar-se d'un TFC tenim alguns requeriments més. Hem de fer una memòria que expliqui el treball que s'ha fet, i una presentació que mostri el més destacable del proces de desenvolupament del TFC i els seus resultats.
L'accés a la informació requerida pel client es fa mitjançant Pentaho BI Platform.

3.3.2.1 Funcionals

Les dades que el client ens aporta mitjançant uns fitxers les hem de passar per un proces d'extracció, transformació i càrrega. A aquest proces se l'anomena ETL (*Extract, Transform and Load*). El resultat d'aquest proces és una base de dades relacional preparada per ser utilitzada per tractar la informació que mostraran els informes. Aquests permeten obtenir i visualitzar aquesta informació dintre d'una temporalitat a nivell d'any i són els següents:

- 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

Tota aquesta informació es podrà consultar de forma agregada, per comarca/província, tipus d'establiment i categoria.

3.3.2.2 No funcionals

El sistema té aquestes propietats que el defineixen.

Seguretat
Pentaho BI Platform disposa d'un sistema de seguretat en l'accés a la informació que serà utilitzat pel portal per permetre aquest accés als usuaris autoritzats a cada un dels informes.

Usuaris	**Funcions**
Bàsic	recopila informació bàsica, principalment informes.
Avançat	pot analitzar les dades lliurement per extreure informació el més valuosa possible, especialment la que es troba analitzant

	les dades.
Administrador	fa les càrregues, resol les incidències, gestiona els usuaris i els permisos d'accés a la informació que tenen els usuaris.

Portabilitat
En tractar-se d'un sistema que presenta la informació en format web, s'ha de disposar d'un navegador web per veure la informació.

Rendiment
Un Magatzem de dades té com a característica que les dades que es consulten s'han de mostrar de forma immediata. Si s'han de mostrar dades calculades han d'estar prèviament calculades o ser càlculs prou ràpids.

Facilitat d'ús
L'usuari bàsic no ha de tenir cap coneixement especial per recopilar informació bàsica, que principalment seran informes. L'usuari avançat ha de tenir coneixements especialitzats perquè ha de poder analitzar les dades lliurement per extreure informació el més valuosa possible, especialment la que es troba analitzant les dades.

Fiabilitat
El sistema està construït usant MySQL i Pentaho. La fiabilitat és la proporcionada per ells i és prou bona com per ser molt usats.
El Centre de Software de Ubuntu diu això sobre MySQL.

> MySQL is a fast, stable and true multi-user, multi-threaded SQL database server. SQL (Structured Query Language) is the most popular database query language in the world. The main goals of MySQL are speed, robustness and ease of use.

A http://www.pentaho.com/partners/technology/ podem veure això:

> Pentaho Technology Partners work with us to ensure sustainable, supportable integration of our products, making a wide range of compatible technologies available to customers. Pentaho's established relationships with these partners enable us to work together to support integration points and resolve joint customer issues.

3.3.3 Model conceptual
De les dades obtingudes dels arxius surten 3 fets; Municipi, Equipament i Establiments. Amb les següents dimensions:

Dimensió	**Fet**		
	Municipi	**Equipament**	**Establiments**
Comarca	☑	☑	☑
Ambit	☑	☑	☑

Provincia	☑	☑	☑
Categoria		☑	☑
Area			☑
Tipus_dada			☑

Cal dir que als informes no es demana l'àmbit, però forma part de la jerarquia d'agregació.
La granularitat es descriu en la jerarquia d'agregació de l'apartat 3.3.5 a on es veu la relació d'agregació que hi ha entre elles.

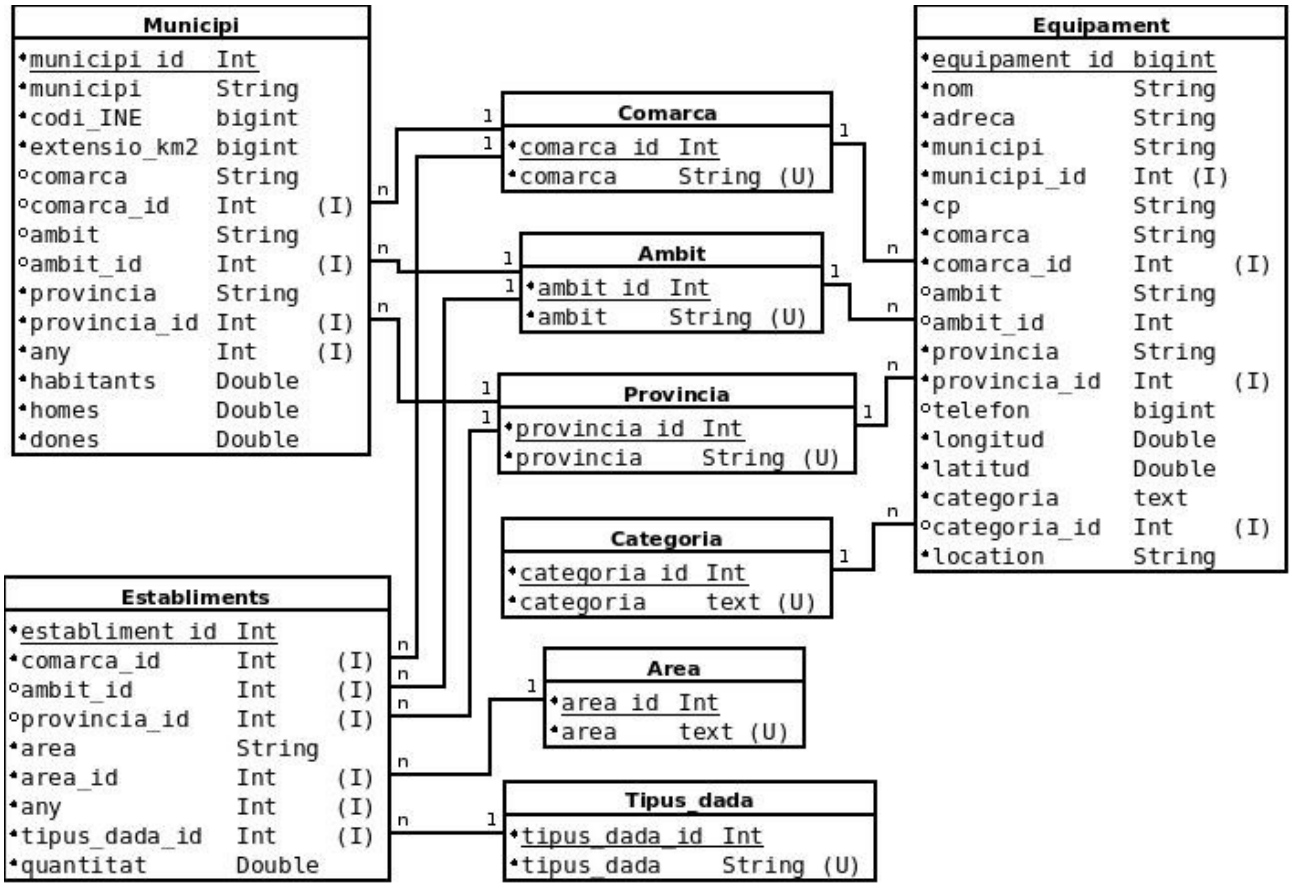
Gränul	**Fet**		
	Municipi	**Equipament**	**Establiments**
Catalunya	☑	☑	☑
Àmbit	☑ (No es demana)	☑ (No es demana)	☑ (No es demana)
Província	☑	☑	☑
Comarca	☑	☑	☑
Municipi	☑ (No es demana)	No es demana	No es demana

La llista dels camps i el seu tipus, per a cada fet i cada dimensió, es pot veure a l'apartat 3.3.4.

Per a la taula Municipi les dades dades son dels anys 2006 a 2012.
Per a la taula Establiments les dades son dels anys 2006 a 2012.
Per a la taula Equipaments les dades no tenen any.

3.3.4 Disseny de la BD / Diagrama E-R

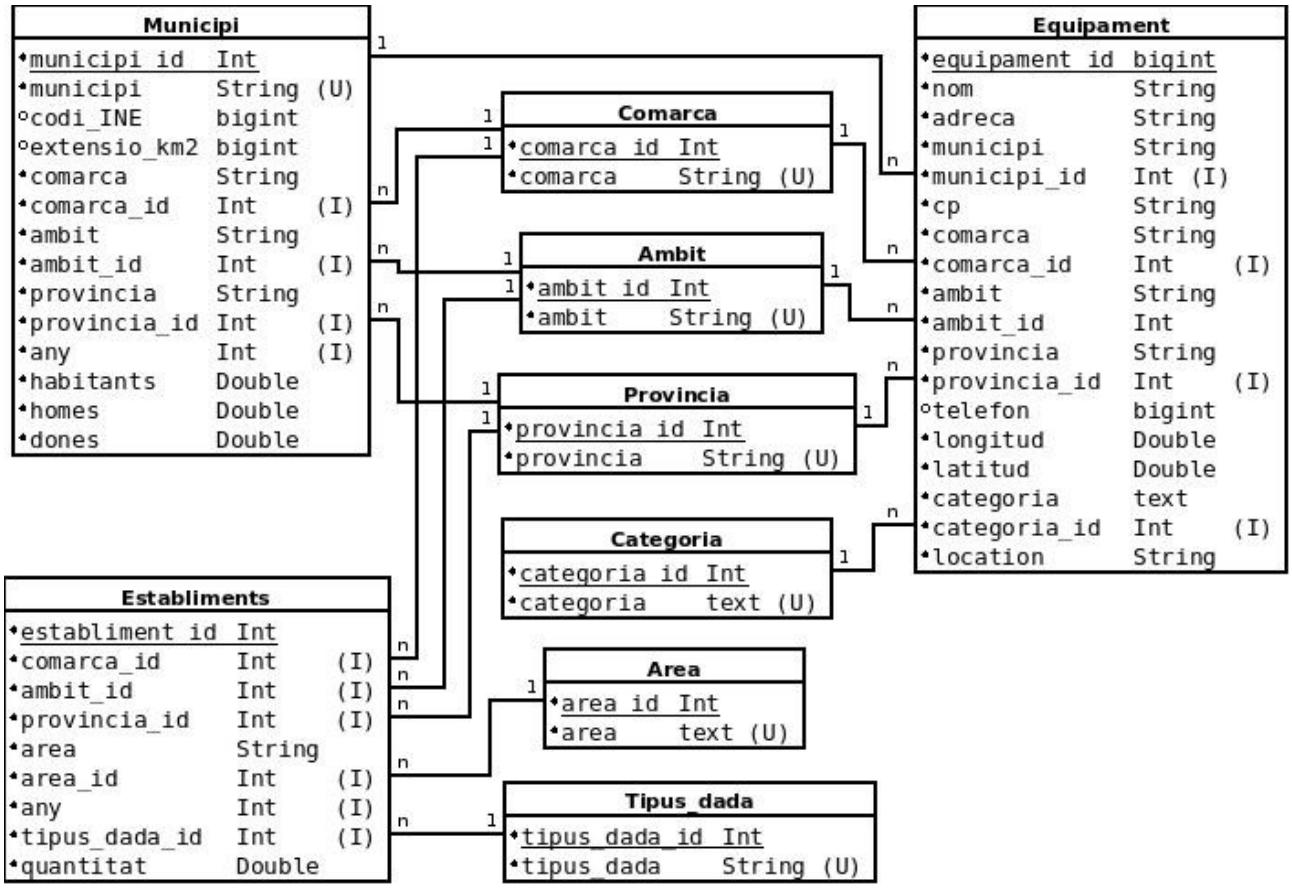
El disseny de la BD ha de ser capaç de suportar els requeriments de dades que els informes demanats necessiten. I facilitar que una sentència SQL sigui aprofitable per obtenir conjunts de dades diferents sense canviar camps, nomes el que han de complir les clàusules *WHERE*. Aquest aprofitament de les sentències SQL obliga a fer canvis en l'estructura de les dades rebudes normalitzant files. A continuació es mostra el disseny, que mes endavant s'explica.



Les taules 'Comarca', 'Ambit' i 'Provincia' tenen la funció de mantenir la integritat referencial entre les dades contingudes a les taules de fets, 'Municipi', 'Equipament' i 'Establiments'. El disseny de les taules s'ha pensat per a que a les consultes, per exemple a la taula Municipi, no s'hagi d'anar a buscar la dada comarca a la Taula Comarca, millorant d'aquesta manera el rendiment.
Les taules 'Comarca', 'Ambit', 'Provincia', 'Categoria', 'Area' i 'Tipus_dada' contenen els identificadors de les taules a les quals estan associats i el text al que fan referencia.
Els camps identificadors, acabats en '_id', serviran per a fer les cerques, aquesta és la raó per la qual son de tipus Index. Els camps Unique asseguren que no hi hagi mes d'un camp identificador associat a un mateix camp de text.
Un dels requeriments és poder obtenir les dades per comarca/província, tipus d'establiment i categoria, per a complir-lo hi han camps identificadors. En el cas de la taula 'Municipi' cal afegir la comarca a la qual pertany cada municipi, que es pot obtenir de la taula 'Equipament' amb "SELECT distinct municipi, comarca FROM dw.equipament order by 1;".
L'àmbit no és necessari per a l'obtenció de dades, però tenint la dada comarca i sabent quines comarques hi ha a cada àmbit es pot incloure. Es podria posar com un camp de comarca, però fent-ho d'aquesta manera s'evita la formació d'un floc de neu, que perjudicaria el rendiment.
La província es pot afegir a partir de la informació del 'codi_INE'.

No s'han d'usar per als informes dades com municipi_id a la taula 'Equipament', i no tindria els mateixos municipis que a la taula 'Municipi' ni els noms estan expressats de la mateixa manera quan comencen per "El" o "La" o "L'", que en el cas de la taula 'Municipi' comença amb majúscula i està entre parèntesi. Es podria fer una taula comú de municipis però s'hauria de normalitzar el nom dels

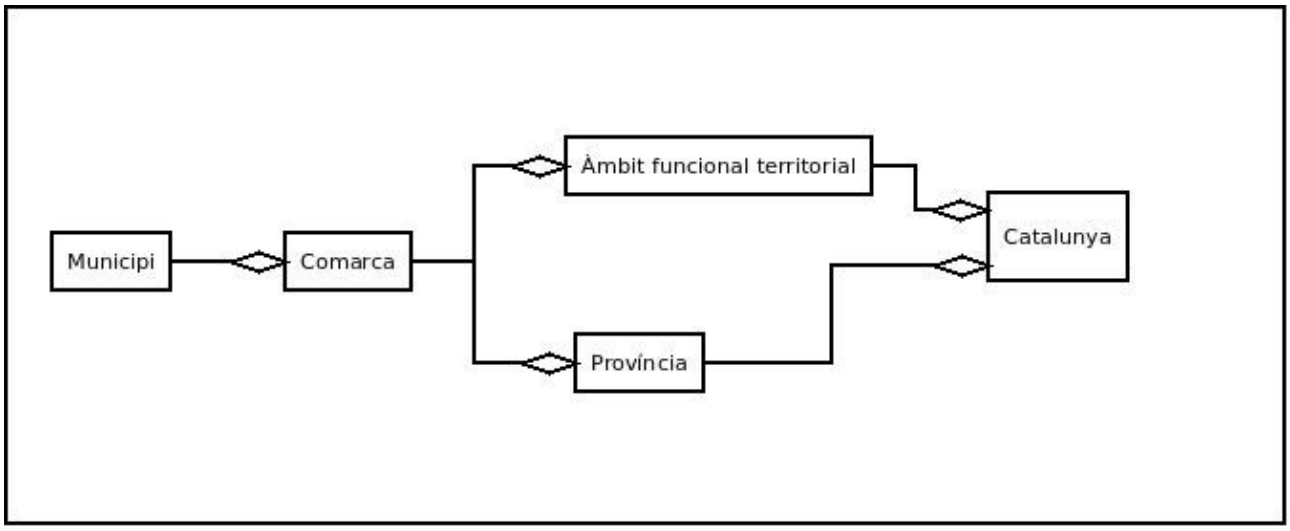
municipis i obtenir dades dels municipis nous provinents de la taula 'Equipament'. Aquesta tasca pot ser un treball futur i la BD quedaria com es veu a continuació.



3.3.5 Model multi-dimensional detallat

Al model multi-dimensional tenim una jerarquia d'agregació que defineix quines dades formen part d'una altra dada.

Diagrama de la jerarquia d'agregació:



Ja s'ha comentat a l'apartat 3.3.4 que a la taula 'Municipi' cal afegir la comarca, l'àmbit i la província. A l'apartat 3.3.7 es detalla quines comarques composen cada àmbit. El municipi no és necessari i no s'implementa. La bifurcació Àmbit funcional territorial / Província no és necessaria ja que no es demana informació per Àmbit funcional territorial.

Els detalls de les dimensions queden reduïts com es veu a l'apartat anterior en haver optat per incloure la informació a la pròpia taula de fet per millorar el rendiment.

3.3.6 Procés ETL a alt nivell

ONdO proporciona uns fitxers de dades que s'han de carregar en una base de dades, per a fer-ho s'usen uns processos ETL en els quals s'han de corregir les dades per garantir la fiabilitat. També s'ha de transformar el format de les dades per a ser útils per a les sentències SQL, en comptes de fer moltes sentències SQL per obtenir les dades, només canviant les condicions a complir de la clàusula WHERE s'han d'obtenir. Aquestes tasques no les pot fer el personal de ONdO, ja que requereixen coneixements de bases de dades i programació.
Una vegada fetes les transformacions, les dades es carreguen a la base de dades MySQL.

A continuació detallo les tasques dels processos ETL.

poblacio.csv
La taula final es dirà 'Municipi'.
Extracció
Utilitzo un pas d'entrada CSV file input per fer l'extracció de les dades de l'arxiu, té delimitador “;” i no té codificació especial. No conté valors nuls.

Transformació
Camps en que estan ordenades les dades.

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Municipi	String		43		€	,	.	ninguno
2	Codi INE	Integer	#	15		€	,	.	ninguno
3	Població 2012	Number	#.#	15		€	.	,	ninguno
4	Població 2011	String		9		€	,	.	ninguno
5	Població 2010	String		9		€	,	.	ninguno
6	Població 2009	String		9		€	,	.	ninguno
7	Població 2008	String		9		€	,	.	ninguno
8	Població 2007	String		9		€	,	.	ninguno
9	Població 2006	Number	#.#	15		€	.	,	ninguno
10	Població 2012 homes	Number	#.#	15		€	.	,	ninguno
11	Població 2012 dones	Number	#.#	15		€	.	,	ninguno
12	Extensió (km2)	Integer	#	15		€	,	.	ninguno

Homogeneïtzació de dades:
Les dades de població per als diferents anys estan en camps de tipus *Number* o *String* i amb diferent format, s'han d'unificar al tipus Number.
S'eliminen els caràcters conflictius als noms de camp, espais i parèntesis.
S'han de passar les dades de binary-string a normal per evitar posteriors problemes.
Les dades de població per al 2012 estan en milers, excepte per a Barcelona que està en milions, la dada podria ser més exacta si es pregués la suma d'homes i dones, per exemple per a Barcelona. Cal multiplicar prèviament la dada de Barcelona per 1000. I després passar-les totes a unitats multiplicant per 1000. Si passes les dades a milers es podria perdre precisió de dades depenent del nombre de decimals. Aquestes dades són de tipus *Number* i ho deixarem com està perquè anirà bé per fer operacions.
A on hi ha la dada “n.d.” s'ha de canviar per 0.
S'ha d'afegir un camp per a 'municipi_id', que s'omplirà amb una seqüència.
Les dades de població per al 2006 per a Barcelona estan en milions, s'han de multiplicar per 1000 com a pas previ. Ara estan en milers excepte per als municipis petits de menys de 1000 habitants, a on estan en unitats, per corregir-ho s'ha de multiplicar per 1000 excepte a on “població 2007 < (població 2006 + població 2006/2)”. D'aquesta manera ja les tenim en unitats.
Les dades de població d'homes per al 2012 i de dones per al 2012 estan en milers, incloent les de Barcelona. Excepte per a les dades de menys d'un miler, a on estan en unitats, per a corregir-ho prèviament les dades del 2012 s'hauran tractat com s'ha indicat, aquestes dues columnes d'homes i dones es multiplicaran per 1000 excepte a on “(població 2012 homes * 1000) > població 2012” o “(població 2012 dones * 1000) > població 2012” segons la columna.

Per a la població dels anys 2011 al 2007 les dades estan en unitats i són de tipus *string*. Cal canviar a Number.
Cal fer una normalització de fila per als camps de població del 2012 al 2006, s'ha de fer per no haver de canviar de sentència SQL per a cada un dels anys que es vulgui consultar. S'obtenen els nous camps 'any' i 'habitants'. El camp 'any' contindrà els noms dels antics camps sense la part de text, i el camp 'habitants' contindrà les dades de població.

<u>Llistat de municipis</u>
El llistat més complert de municipis està a l'arxiu d'equipaments, aquest llistat ens serviria per afegir un camp id a la taula de municipis en el treball futur que s'ha comentat prèviament a l'apartat 3.3.4.

```
// els municipis que estan a equipaments i no a poblacio
SELECT distinct lower(dw.equipament.municipi)
FROM dw.equipament
```

```
WHERE lower(dw.equipament.municipi) not in (SELECT replace(replace(lower(dw.municipi_de_poblacio.municipi),'(',' '),')','') FROM dw.municipi_de_poblacio)
ORDER BY 1;


// els municipis que estan a poblacio i no a equipaments
SELECT distinct replace(replace(lower(dw.municipi_de_poblacio.municipi),'(',' '),')','')
FROM dw.municipi_de_poblacio
WHERE replace(replace(lower(dw.municipi_de_poblacio.municipi),'(',' '),')','') not in (SELECT lower(dw.equipament.municipi) FROM dw.equipament)
ORDER BY 1;
```

En el resultat d'aquest SQL veiem que:
Cabrera d'Anoia provinent de poblacio, no existeix
cruïlles, monells i sant sadurní de l'heura és correcte , s'ha de corregir a equipaments, a on està sense “,”.
Vimbodí i Poblet no és un municipi

Conclusió: la llista de municipis d'equipaments és més complerta.

Càrrega
La càrrega es fa sobre la taula 'Municipi' de la BD 'dw' de MySQL.

Equipaments.csv
Per a poder transformar el gran nombre de files que conte aquest arxiu s'ha de de fer aquest canvi a la configuració de PDI -> Editar -> Configuración -> Miscelaneos -> Nr de filas = 50000

D'aquest arxiu es pot extreure la llista de municipis i comarques a les que aquests municipis pertanyen.

Extracció
Utilitzo un pas d'entrada CSV file input per fer l'extracció de les dades de l'arxiu, té separador “,” i codificació UTF-8.

Transformació
Camps en que estan ordenades les dades.

#.	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	nom	String		145		€	,	.	ninguno
2	adreca	String		112		€	,	.	ninguno
3	municipi	String		43		€	,	.	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		382		€	,	.	ninguno
10	Location	String		83		€	,	.	ninguno

Homogeneïtzació de dades:
S'han de revisar les dades, als camps latitud i longitud s'ha de controlar els que estiguin fora del rang per a Catalunya. Per a alguns es veu el punt decimal mal posat.
El telèfon està incomplet o és null a molts equipaments, però no els usarem.
A partir d'aquest arxiu podem tenir la comarca a la qual pertany cada municipi i la província de cada comarca. Hi ha alguns municipis que estan a població i no estan a equipaments, d'aquests es pot buscar manualment. Es pot usar:
SELECT Distinct municipi, cp, comarca FROM dw.equipament where cp <> "-" group by 1 order by 1;
I la província a la qual pertany cada comarca:
SELECT Distinct municipi, cp, comarca FROM dw.equipament where cp <> "-" group by 3 order by 3;
Cal afegir l'àmbit i la província. També el camp identificador de fila 'equipament_id' amb una seqüència. I la resta d'identificadors.

Càrrega
La càrrega es fa sobre la taula 'Equipament' de la BD 'dw' de MySQL.

establiments 2006.txt ... ...establiments 2012.txt
Extracció
Utilitzo un pas d'entrada CSV file input per fer l'extracció de les dades de l'arxiu, té delimitador “;” i no té codificació especial. No conté valors nuls.

Transformació
Camps en que estan ordenades les dades.

#.	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Field_000	String		49		€	,	.	ninguno
2	Field_001	String		12		€	,	.	ninguno
3	Field_002	String		8		€	,	.	ninguno
4	Field_003	String		7		€	,	.	ninguno
5	Field_004	String		5		€	,	.	ninguno
6	Field_005	String		12		€	,	.	ninguno
7	Field_006	String		8		€	,	.	ninguno
8	Field_007	String		7		€	,	.	ninguno
9	Field_008	String		7		€	,	.	ninguno
10	Field_009	Date	yyyy/MM/dd HH:mm:ss.SSS			€	,	.	ninguno
11	Field_010	String		14		€	,	.	ninguno
12	Field_011	String		9		€	,	.	ninguno
13	Field_012	String		5		€	,	.	ninguno
14	Field_013	String		14		€	,	.	ninguno
15	Field_014	String		9		€	,	.	ninguno
16	Field_015	String		7		€	,	.	ninguno
17	Field_016	Date	yyyy/MM/dd HH:mm:ss.SSS			€	,	.	ninguno
18	Field_017	String		12		€	,	.	ninguno
19	Field_018	Integer	#	15		€	,	.	ninguno
20	Field_019	Integer	#	15		€	,	.	ninguno
21	Field_020	String		11		€	,	.	ninguno
22	Field_021	String		5		€	,	.	ninguno
23	Field_022	String		7		€	,	.	ninguno
24	Field_023	Number	#.#	15		€	.	,	ninguno
25	Field_024	Number	#.#	15		€	.	,	ninguno
26	Field_025	String		11		€	,	.	ninguno
27	Field_026	String		7		€	,	.	ninguno
28	Field_027	Date	yyyy/MM/dd HH:mm:ss.SSS			€	,	.	ninguno
29	Field_028	String		12		€	,	.	ninguno
30	Field_029	String		5		€	,	.	ninguno
31	Field_030	String		7		€	,	.	ninguno
32	Field_031	String		5		€	,	.	ninguno
33	Field_032	String		11		€	,	.	ninguno
34	Field_033	String		5		€	,	.	ninguno
35	Field_034	String		7		€	,	.	ninguno
36	Field_035	String		6		€	,	.	ninguno

Canvi del nom dels camps.

#.	Nombre campo	Renombrar a
1	Field_000	area
2	Field_001	n_hotels
3	Field_002	n_campings
4	Field_003	n_turisme_rural
5	Field_004	n_total
6	Field_005	p_hotels
7	Field_006	p_campings
8	Field_007	p_turisme_rural
9	Field_008	p_total
10	Field_009	
11	Field_010	n_hotels_estrelles_or
12	Field_011	n_hotels_estrelles_argent
13	Field_012	n_hotels_total_estrelles
14	Field_013	p_hotels_estrelles_or
15	Field_014	p_hotels_estrelles_argent
16	Field_015	p_hotels_total_estrelles
17	Field_016	
18	Field_017	n_campings_cat_1
19	Field_018	n_campings_cat_2
20	Field_019	n_campings_cat_3
21	Field_020	n_campings_cat_privat
22	Field_021	n_campings_total_cat
23	Field_022	p_campings_cat_1
24	Field_023	p_campings_cat_2
25	Field_024	p_campings_cat_3
26	Field_025	p_campings_cat_privat
27	Field_026	p_campings_total_cat
28	Field_027	
29	Field_028	n_turisme_rural_allotjament_independent
30	Field_029	n_turisme_rural_masia
31	Field_030	n_turisme_rural_casa_poble
32	Field_031	n_turisme_rural_total
33	Field_032	p_turisme_rural_allotjament_independent
34	Field_033	p_turisme_rural_masia
35	Field_034	p_turisme_rural_casa_poble
36	Field_035	p_turisme_rural_total

Tenim columnes redundants:
n_hotels = n_hotels_total_estrelles
n_campings = n_campings_total_cat
n_turisme_rural = n_turisme_rural_total
p_hotels = p_hotels_total_estrelles
p_campings = p_campings_total_cat
p_turisme_rural = p_turisme_rural_total

Eliminació de camps buits.

Campos a eliminar :

#.	Nombre campo
1	Field_009
2	Field_016
3	Field_027

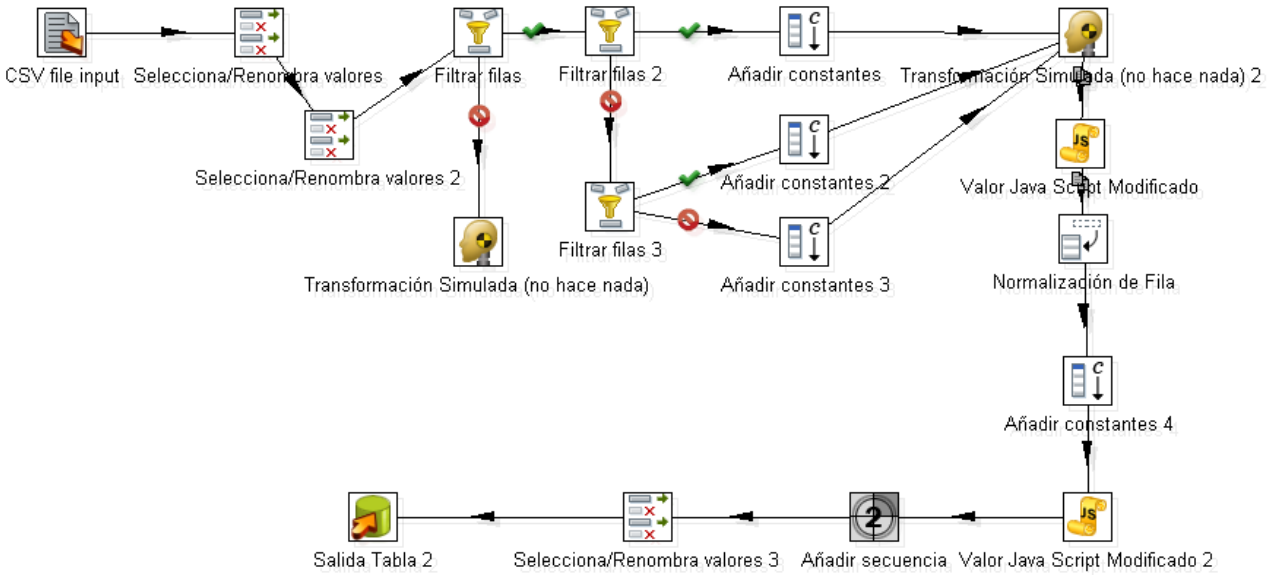
Les 10 primeres files no són de dades, són de comentaris i noms de columnes, es poden eliminar deixant només les que compleixin “n_hotels IS NOT NULL AND n_campings IS NOT NULL”.

A les files que compleixin que el camp area sigui "Catalunya" o "Metropolità" o "Comarques Gironines" o "Camp de Tarragona" o "Terres de l'Ebre" o "Ponent" o "Comarques Centrals" o "Alt

Pirineu i Aran" se li afegeix el camp tipus_area = 0, 0 correspon a àmbit. Si el camp area és "Barcelona" o "Girona" o "Lleida" o "Tarragona" se li afegeix el camp tipus_area = 1, corresponent a província. En cas contrari se li afegeix el camp tipus_area = 2, corresponent a comarca. A totes les files se li afegeix el camp any amb el contingut 2006.
El tipus d'àrea pot ser: 0 per àmbit, 1 per província o 2 per comarca.
Els camps String de dades numèriques s'han de passar a Number.
Es normalitza fila per als camps referits a nombre i places d'establiments, els noms de les columnes passen al camp tipus_dada i el contingut al camp quantitat.
S'afegeixen els camps id que falten.

Càrrega
La càrrega es fa sobre la taula 'Establiments' de la BD 'dw' de MySQL.

Passos del proces ETL.



Per a la resta d'anys el proces és similar, a l'últim any tenim dades mes avall de les files habituals.

3.3.7 Tractament d'errors en la càrrega (qualitat de les dades)

Poblacions
A l'hora d'obtenir un id_poblacio, s'observa que hi ha diferència en el nom per a una mateixa població (“Prat de Llobregat el” (a l'arxiu equipaments), “Prat de Llobregat (El)” (a l'arxiu poblacio)). La diferència en el nom es pot corregir per unificar els noms. La llista més completa està a equipaments.

Si tenim una única llista de municipis, als municipis que estiguin a equipaments però no a població no tindrem dades d'habitants, per tant, per evitar errors de càlculs sense dades s'ha de comprovar que els càlculs es facin amb dades. D'aquesta manera podem tenir una dimensió conformada de municipis que s'usi a més d'una estrella. Si estigues completa permetria canviar el tema objecte

d'anàlisi (Drill-across) per navegar d'una estrella a una altra.

L'àmbit s'ha d'obtenir a partir de la informació de quines comarques composen cada àmbit, es detallaran més endavant.

El codi_INE es pot usar per saber a quina província pertany una població, els municipis que provenen de la llista d'equipaments no tenen aquest codi. Per a les diferents províncies els codis són:

Barcelona	**Girona**	**Lleida**	**Tarragona**
8000-8999	17000-17999	25000-25999	43000-43999

Aquest codi no és el mateix que el codi postal, per exemple Lleida té el codi_INE 25120, però el codi postal 25120 pertany a Alfarras segons http://www.correos.es

Els codis INE es poden trobar a
http://www.ine.es/daco/daco42/codmun/codmun11/11codmunmapa.htm
però no ens aporta res, perquè dels municipis provinents d'equipaments podem extreure la província de la població a través del codi postal.

Els formats impedeixen que veiem com estan realment les dades, per a alguns camps hi ha zeros al començament. Els decimals fan que les dades estiguin expressades en unitats, milers i milions, s'ha d'unificar a unitats.

Les dades de població en una columna per a cada any s'han de posar en una sola columna any per no haver de canviar l'SQL per a cada any.

El municipi de La Canonja té dades de població “n.d.” i zero.
Segons http://ca.wikipedia.org/wiki/Canonja
La Canonja es va constituir com a entitat municipal descentralitzada el 1982. I el Parlament de Catalunya aprovà el 15 d'abril de 2010 la creació del municipi de la Canonja. Per tant la dada que tenim per a l'any 2006 = 0, té el mateix valor que n.d. per a l'any 2007, ja que no hi ha diferència en la situació de la població per als anys 2006 i 2007.

Equipaments
Als camps latitud i longitud s'ha de controlar els que estiguin fora del rang per a Catalunya. Per a alguns es veu el punt decimal mal posat.
El telèfon està incomplet o és null a molts equipaments, però no els usarem.

Establiments
Tenim el següent advertiment:
“*A partir de l'any 2011, a l'article 49 del Decret 183/2010, de 23 de novembre, d'establiments d'allotjament turístic, la capacitat d'allotjament d'un càmping en nombre de places s'obté multiplicant per tres el nombre total d'unitats d'acampada. Aquest canvi comporta un trencament de la sèrie que n'impedeix la comparació interanual, ja que per als anys anteriors el nombre de places d'un càmping s'obtenia multiplicant per dos i mig el nombre total d'unitats d'acampada.*”

A les dades d'establiments tenim els següents àmbits: Metropolità, Comarques Gironines, Camp de Tarragona, Terres de l'Ebre, Ponent, Comarques Centrals, Alt Pirineu i Aran i Penedès. Ens falta saber quines comarques pertanyen a cada àmbit.

Segons http://ca.wikipedia.org/wiki/%C3%80mbit_funcional_territorial tenim els següents àmbits funcionals territorials:

- *Alt Pirineu i Aran: Alta Ribagorça, Alt Urgell, Cerdanya, Pallars Jussà, Pallars Sobirà i Vall d'Aran.*
- *Àmbit metropolità de Barcelona: Baix Llobregat, Barcelonès, Maresme, Vallès Oriental i Vallès Occidental.*
- *Camp de Tarragona: Tarragonès, Alt Camp, Baix Camp, Conca de Barberà i Priorat.*
- *Comarques gironines: Alt Empordà, Baix Empordà, Garrotxa, Gironès, Pla de l'Estany, La Selva i Ripollès.*
- *Comarques Centrals: Bages, Berguedà, Osona, Solsonès i els municipis de la comarca de l'Anoia que ho sol·licitin.*
- *Penedès: Alt Penedès, Baix Penedès, Garraf i Anoia, excepte els municipis que sol·licitin restar adscrits a Comarques Centrals (possiblement 7 municipis de l'Alta Segarra).*
- *Ponent: Garrigues, Noguera, Segarra, Segrià, Pla d'Urgell i Urgell.*
- *Terres de l'Ebre: Baix Ebre, Montsià, Ribera d'Ebre i Terra Alta.*

Llista de comarques de cada província (Font: http://ca.wikipedia.org/wiki/Prov%C3%ADncies_de_Catalunya):

Barcelona	**Girona**	**Lleida**	**Tarragona**
Alt Penedès Anoia Bages Baix Llobregat Barcelonès Berguedà, excepte Gósol Garraf Maresme Osona (excepte els municipis d'Espinelves, Vidrà i Viladrau) Vallès Occidental Vallès Oriental	Alt Empordà Baix Empordà Cerdanya Garrotxa Gironès Pla de l'Estany Ripollès Selva (excepte el municipi de Fogars de la Selva)	Alta Ribagorça Alt Urgell Garrigues Noguera Pallars Jussà Pallars Sobirà Pla d'Urgell Segarra Segrià Solsonès Urgell Vall d'Aran	Alt Camp Baix Camp Baix Ebre Baix Penedès Conca de Barberà Ribera d'Ebre Montsià Priorat Tarragonès Terra Alta

També la podem extreure de l'arxiu equipaments usant els camps cp i comarca.

3.3.8 Automatització procés ETL

En el nostre cas farem una única càrrega de les dades, per tant no cal programar un '*job*' per a una data i hora determinades.

3.4 Implementació

3.4.1 Explicació sobre com s'ha dut a terme el treball demanat

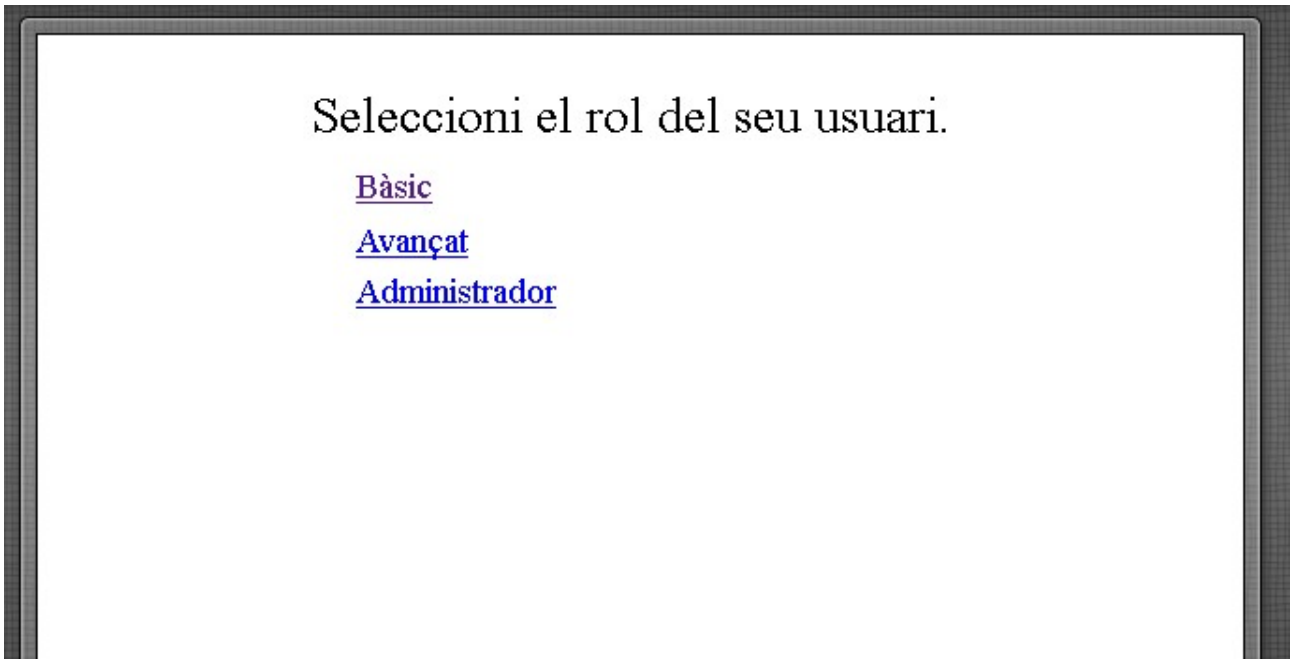
Per a fer el treball demanat he usat PDI (*Pentaho Data Inegration*) per a extreure la informació aportada per ONdO dels arxius, corregir els errors, transformar-la per aconseguir organitzar-la en les taules del magatzem de dades i carregar-la a la base de dades MySQL. A PDI hi ha els *scripts* de creació de taules.

Una altra part ha estat la creació dels informes, que s'alimenten de les dades proporcionades per rutines PL/SQL. Els informes usen paràmetres per evitar haver de fer moltes consultes diferents i molts informes. He creat usuaris amb la consola d'administració i he donat permís a aquests usuaris per veure els informes.

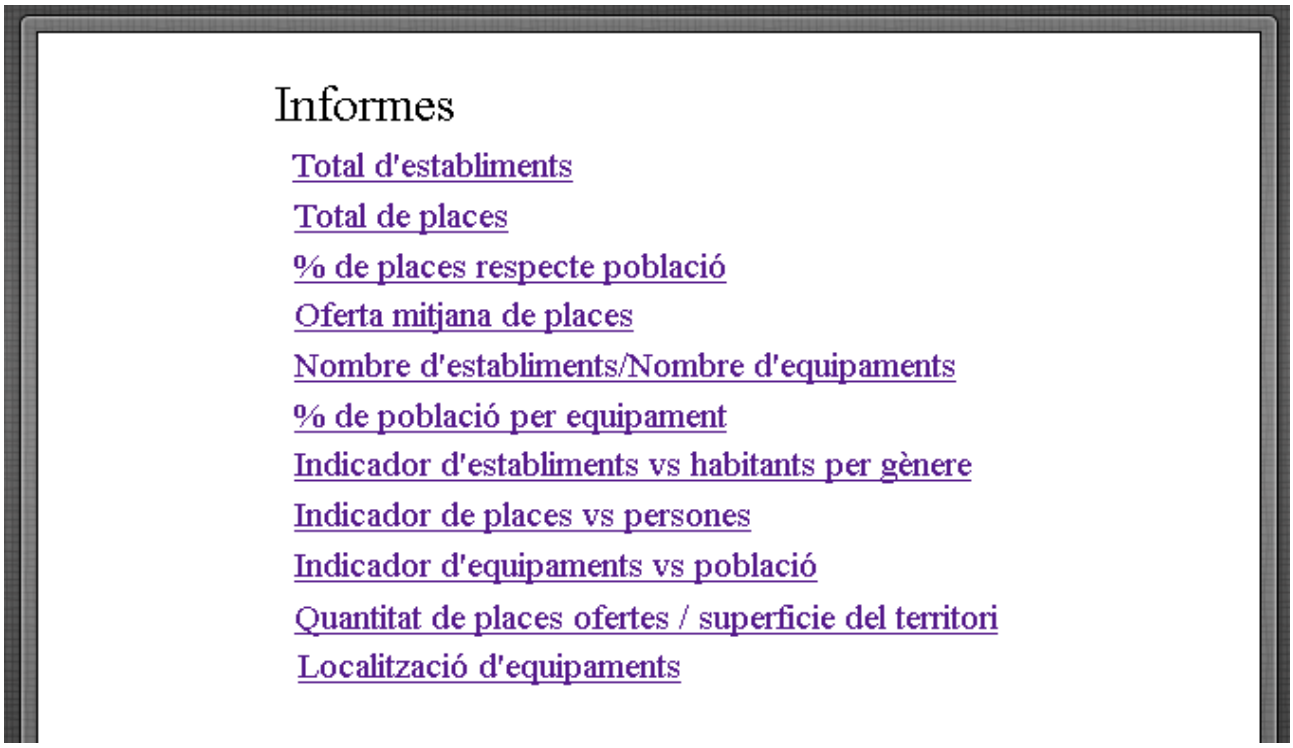
3.4.2 Informes realitzats

He treballat en un sistema de control d'accés en format web que aprofita el *login* de Pentaho. Es selecciona el rol d'usuari, es fa el *login* i es mostra a cada tipus d'usuari els informes als quals pot accedir, malauradament aquests informes perden la part dels paràmetres quan s'accedeix d'aquesta manera. He arribat a la conclusió de que és degut a que despès de fer el login els enllaços dels informes el que fan és canviar la pàgina a on està l'enllaç, però la part dels paràmetres, que està fora de la pàgina, no la pot canviar. La forma de veure els informes amb tota la funcionalitat és amb Pentaho BI Platform.

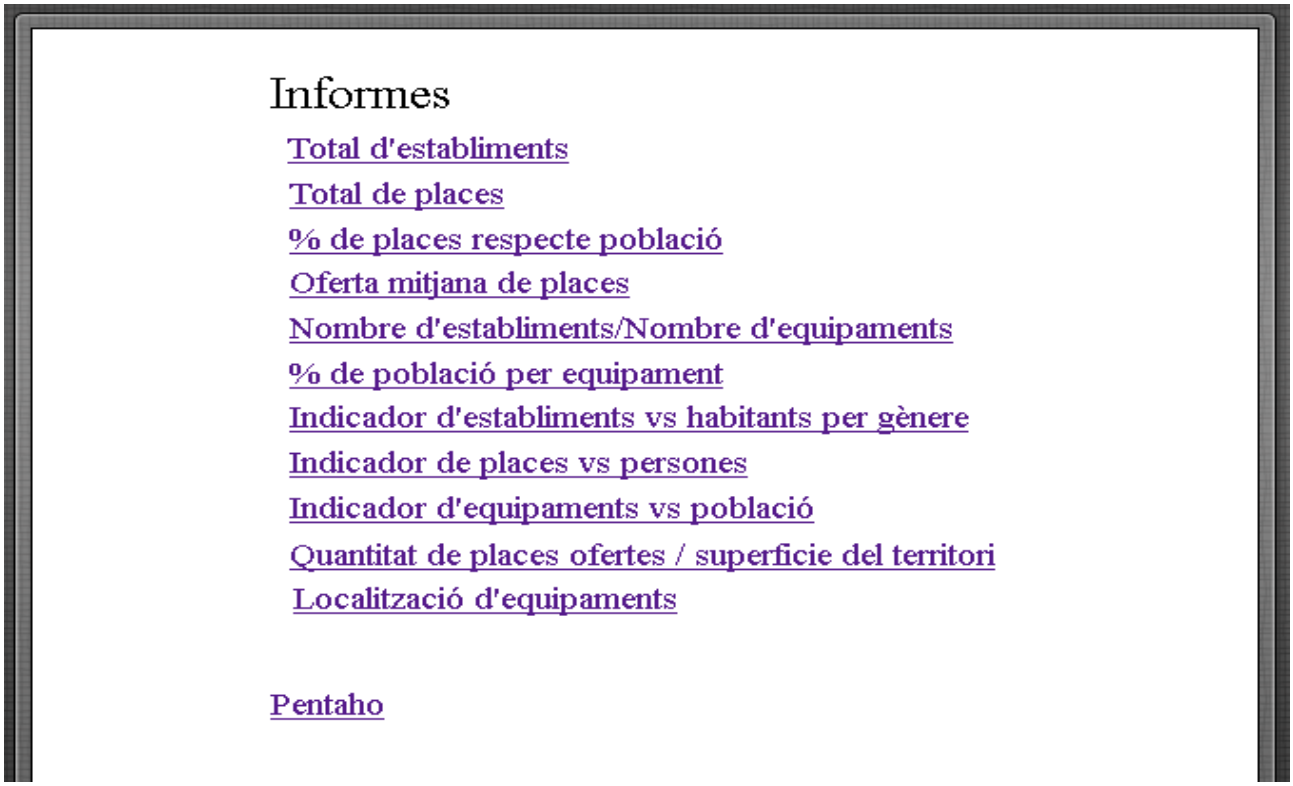
Selecció del rol d'usuari



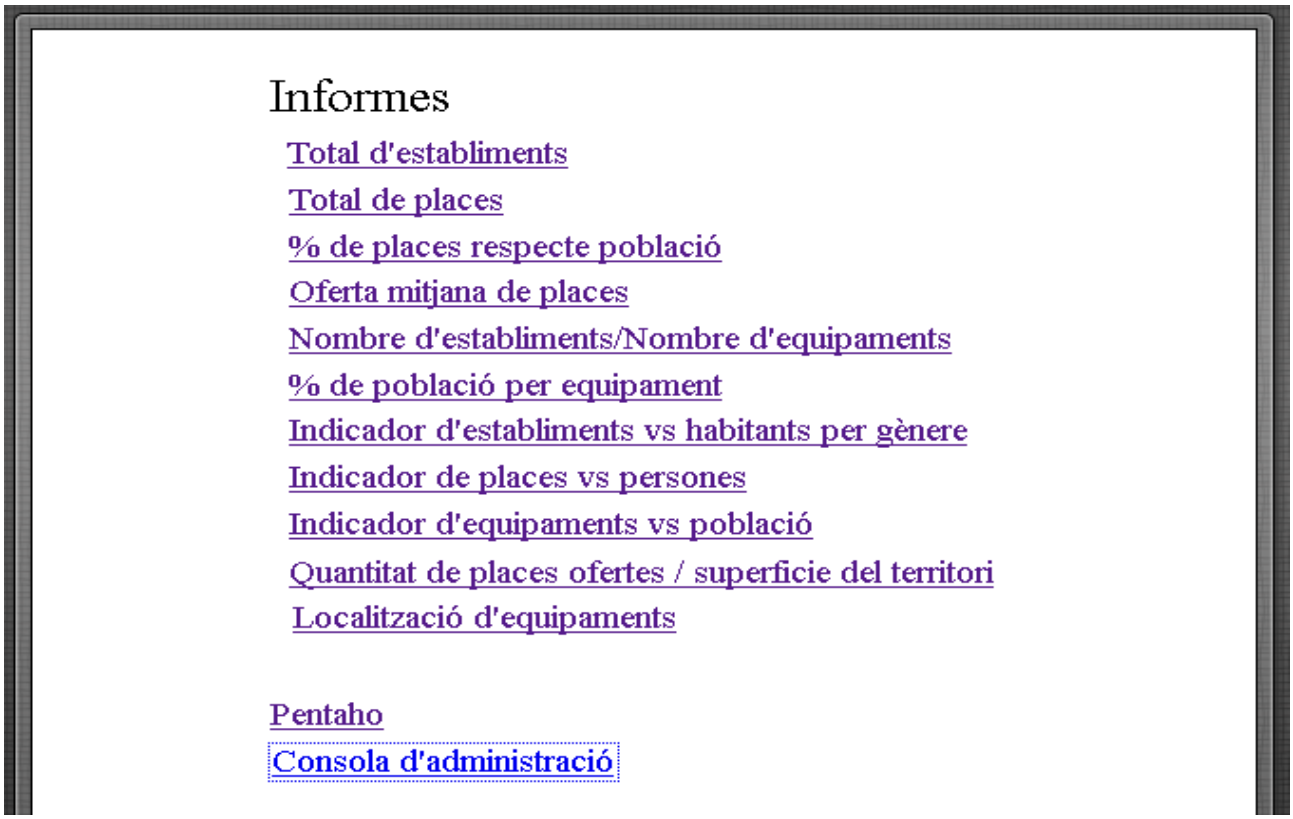
Opcions per a l'usuari bàsic



Opcions per a l'usuari avançat



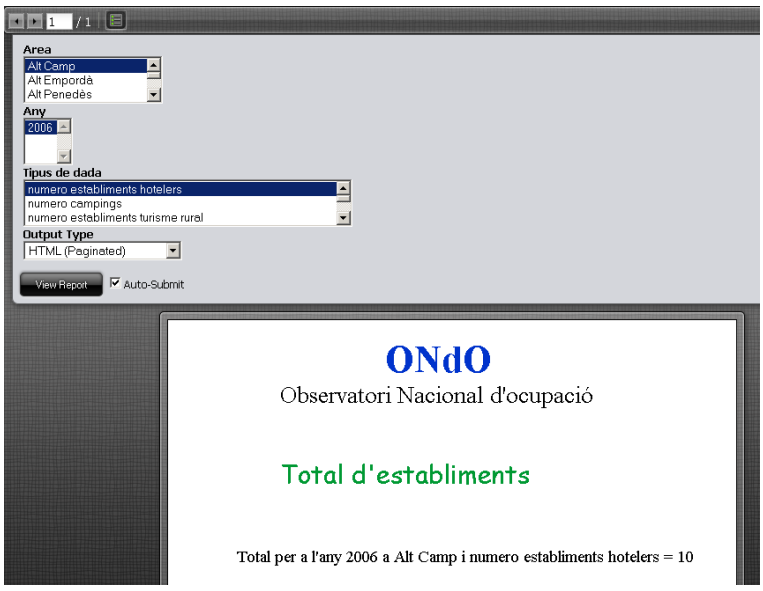
Opcions per a l'usuari administrador



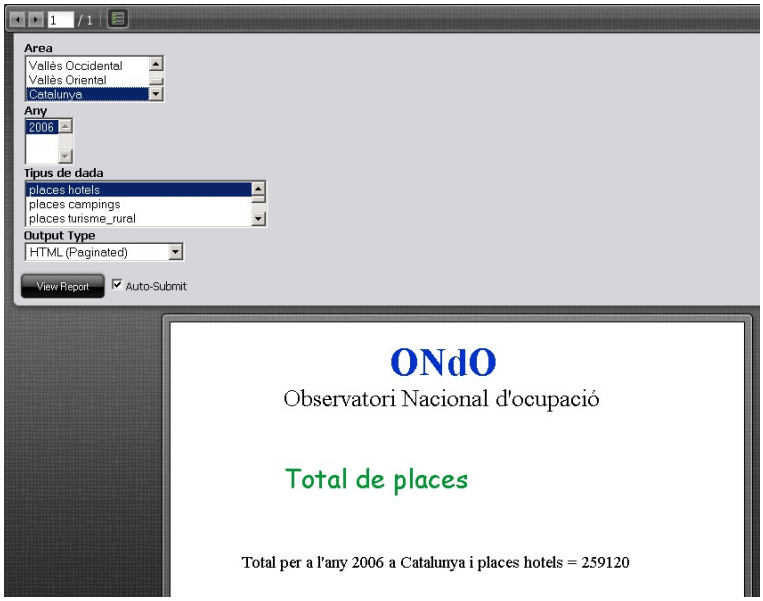
Informes

A continuació es presenten les captures de pantalla dels informes. Es pot observar a cada una els paràmetres que s'usen per a mostrar les dades.

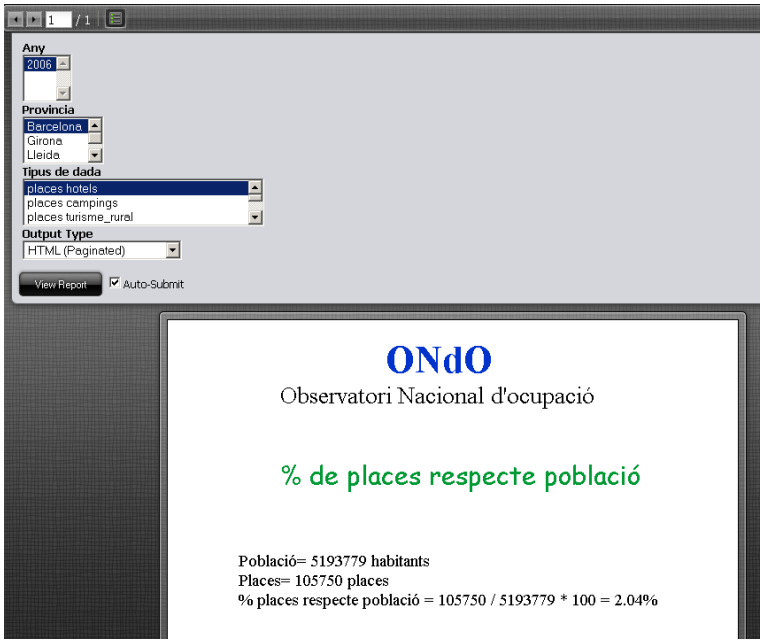
Total d'establiments



Total de places

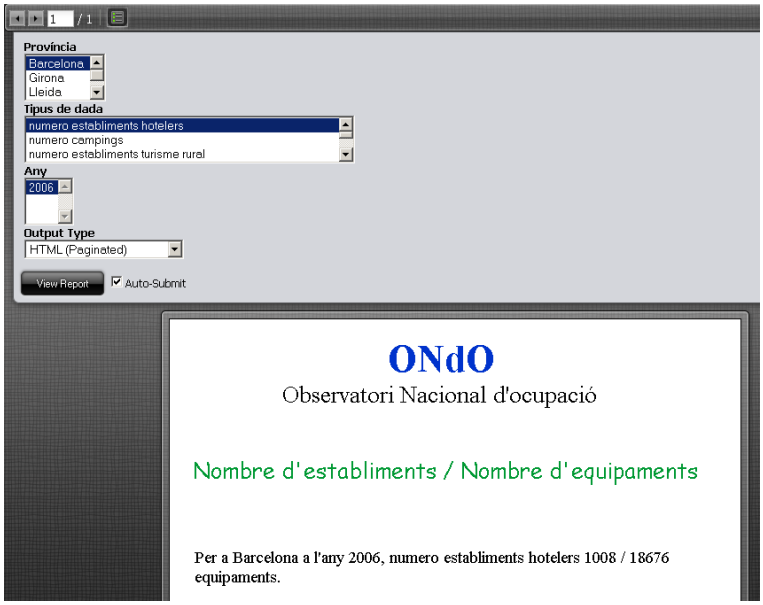


% de places respecte població

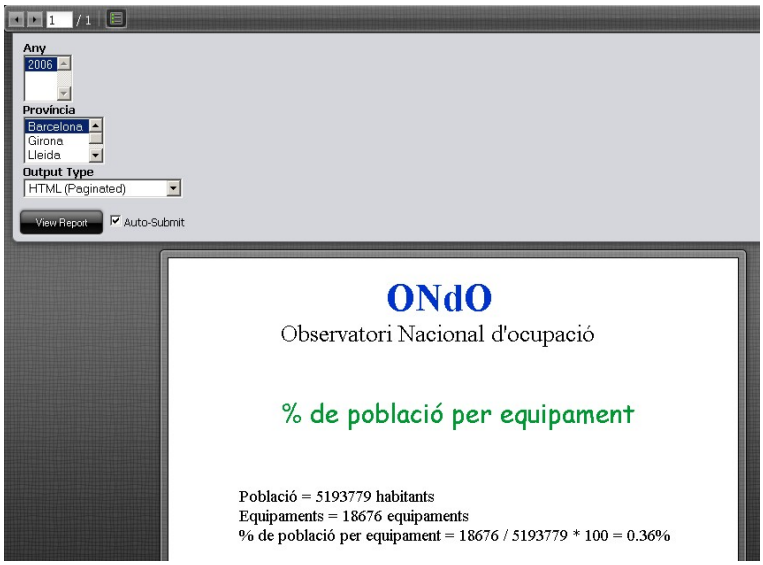


Oferta mitjana de places

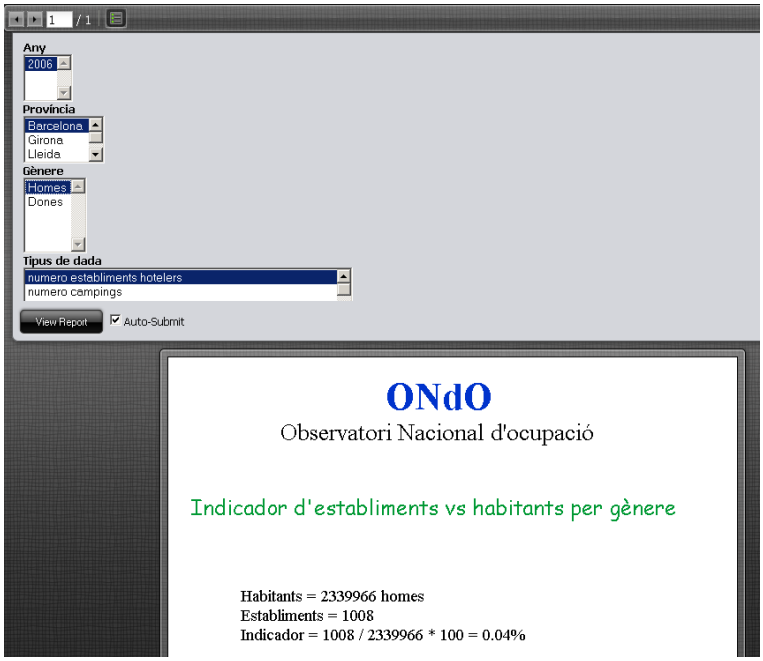
Nombre d'establiments/Nombre d'equipaments



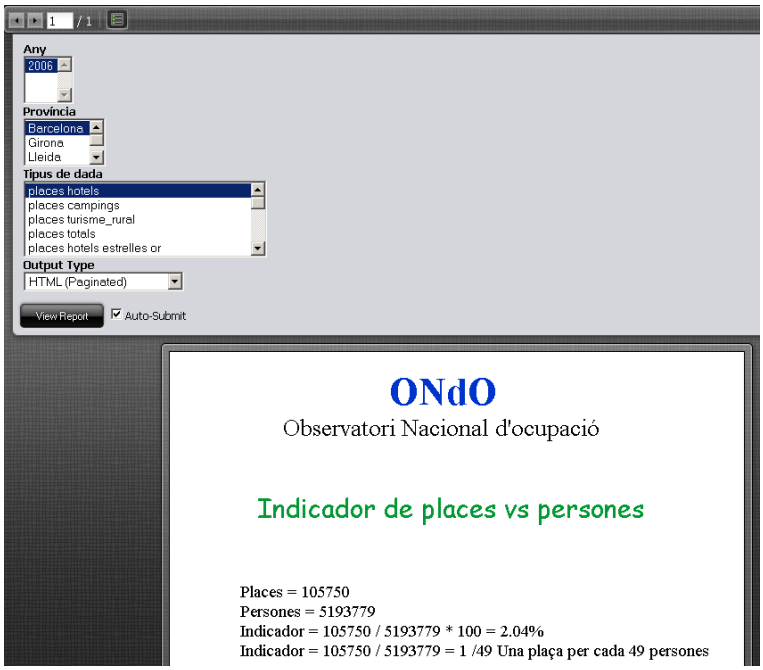
% de població per equipament



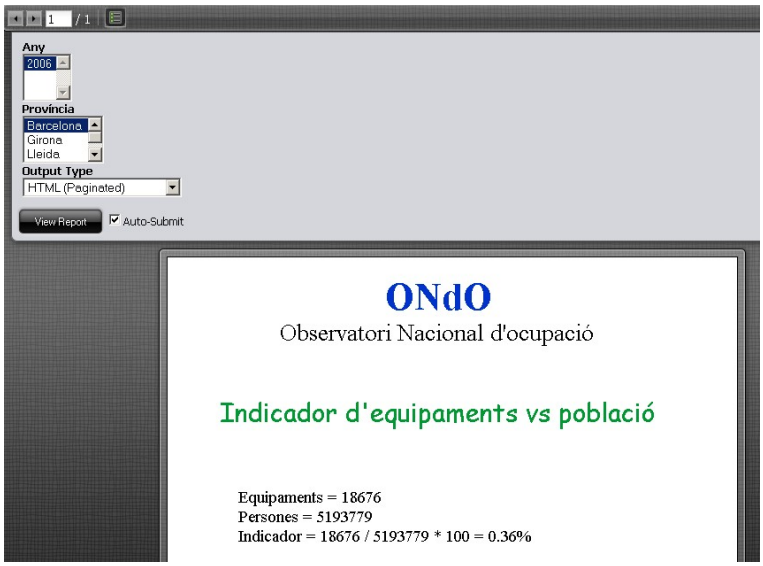
Indicador d'establiments vs habitants per gènere



Indicador de places vs persones



Indicador d'equipaments vs població



Quantitat de places ofertes / superfície del territori

Localització d'equipaments

Aquest informe dona la possibilitat de mostrar sobre un mapa la localització dels equipaments.
Per a fer-ho he usat la pàgina web www.openstreetmap.org, que permet incloure els paràmetres de longitud i latitud.

Treball realitzat sobre la base de dades; taules i rutines.

Les rutines implementades són funcions PL/SQL que tornen les dades requerides tenint en compte els paràmetres seleccionats per a cada informe. Primer es seleccionen els paràmetres, aquests s'envien a la funció PL/SQL i aquesta retorna les dades que usa l'informe per a fer la visualització.

3.4.3 Màquina virtual comprimida

A part de l'usuari administrador donat he creat aquests amb la consola d'administració.

Usuari	**Contrasenya**
basic1	basic1
avançat1	avançat1
administrador1	administrador1

3.4.4 Comentaris rellevants sobre el desenvolupament dut a terme

Rutines PL/SQL d'extracció d'informació.

Informació requerida	**Funció**
Total d’establiments	total_establiments
Total de places	total_places
% de places respecte població	places_respecte_poblacio
Oferta mitjana de places	oferta_mitjana_places

Nombre d'establiments/Nombre d'equipaments	n_establiments_n_equipaments
% de població per equipament	poblacio_per_equipament
Indicador d'establiments vs habitants per gènere	establiments_habitants
Indicador de places vs persones	places_persones
Indicador d'equipaments vs població	equipaments_poblacio
Quantitat de places ofertes / superfície del territori	places_superficie

Establiments
Dels establiments em surten les taules Area, Tipus_dada, Ambit, Comarca i Provincia.
Area

area_id	**area**
1	Alt Camp
2	Alt Empordà
3	Alt Penedès
4	Alt Urgell
5	Alta Ribagorça
6	Anoia
7	Bages
8	Baix Camp
9	Baix Ebre
10	Baix Empordà
11	Baix Llobregat
12	Baix Penedès
13	Barcelonès
14	Berguedà
15	Cerdanya
16	Conca de Barberà
17	Garraf
18	Garrigues
19	Garrotxa
20	Gironès
21	Maresme
22	Montsià

23	Noguera
24	Osona
25	Pallars Jussà
26	Pallars Sobirà
27	Pla d'Urgell
28	Pla de l'Estany
29	Priorat
30	Ribera d'Ebre
31	Ripollès
32	Segarra
33	Segrià
34	Selva
35	Solsonès
36	Tarragonès
37	Terra Alta
37	Urgell
39	Val d'Aran
40	Vallès Occidental
41	Vallès Oriental
42	Catalunya
43	Metropolità
44	Comarques Gironines
45	Camp de Tarragona
46	Terres de l'Ebre
47	Ponent
48	Comarques Centrals
49	Alt Pirineu i Aran
50	Barcelona
51	Girona
52	Lleida
53	Tarragona

Àmbit
SELECT distinct area FROM dw.establiments where tipus_area_id = 0;

ambit_id	ambit
1	Catalunya
2	Metropolità
3	Comarques Gironines
4	Camp de Tarragona
5	Terres de l'Ebre
6	Ponent
7	Comarques Centrals
8	Alt Pirineu i Aran

Comarques
SELECT distinct `establiments`.`area`
FROM `dw`.`establiments` where `establiments`.`tipus_area_id` = 2;

Taula Equipament.
Sql per a la orientació en la correcció de les dades de longitud i latitud.
-- Correcció longitud, per longitud massa petita
SELECT longitud, latitud, municipi, equipament_id FROM dw.equipament order by 1 asc limit 20;

-- Correcció longitud, per longitud massa gran
SELECT longitud, latitud, municipi, equipament_id FROM dw.equipament order by 1 desc limit 20;

-- Correcció latitud, per latitud massa petita
SELECT longitud, latitud, municipi, equipament_id FROM dw.equipament order by 2 asc limit 20;

-- Correcció latitud, per latitud massa gran
SELECT longitud, latitud, municipi, equipament_id FROM dw.equipament order by 2 desc limit 20;

Donat que no disposo de les dades correctes, les deixo tal qual, d'aquesta manera per a les que són incorrectes és més evident que ho són que si intentés fer una aproximació.

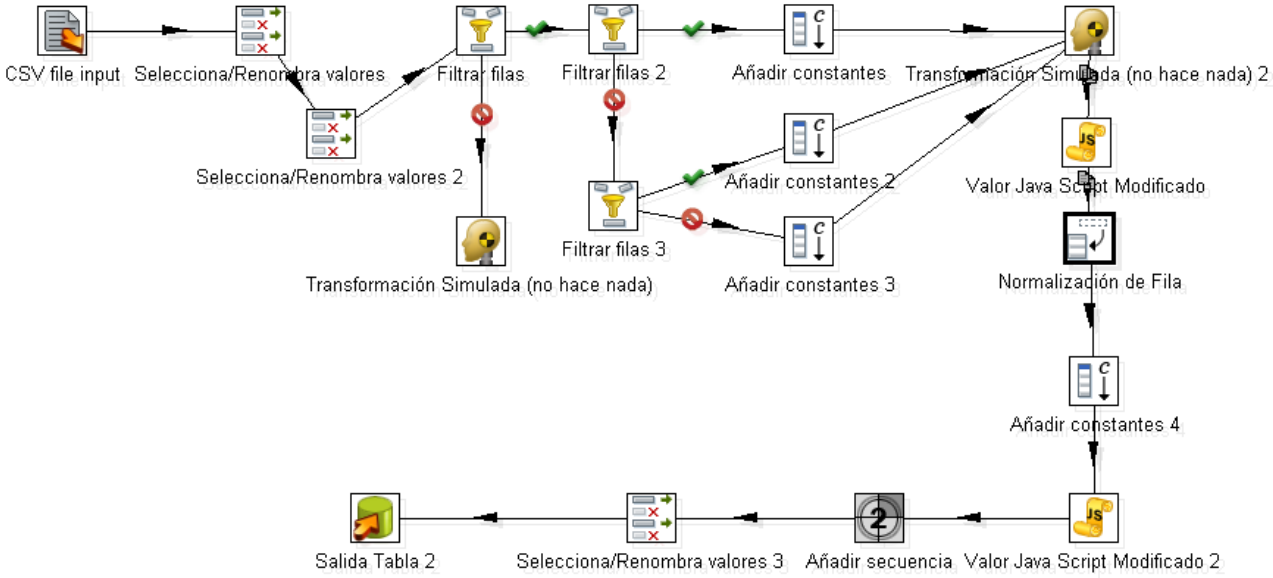
3.4.5 Suposicions fetes, en cas que s'escaigui

Les dades dels equipaments no contenen l'any, ni tan sols de la posada en servei de l'equipament, seria útil també l'any de fi de servei. Això obliga a considerar els equipaments en servei durant tots el anys.

Si el volum de dades fos major o es treballés amb una màquina amb pocs recursos, seria útil fer precàlculs i emmagatzemar-los en taules de la població, de la superfície i del nombre d'equipaments per província i per comarca, això acceleraria les consultes.

3.4.6 Justificació del compliment dels requisits funcionals

He treballat en tots els informes demanats.
Les dades les he hagut d'extreure, transformar i carregar amb PDI (*Pentaho Data Integration*).



Per a transformar les dades ha estat necessari conèixer l'eina i les seves funcions, una vegada fet això ha estat possible fer els processos ETL. Per a les dades dels establiments he usat una funció de normalització de taula que després m'ha permès fer sentències SQL aprofitables per a tots els tipus d'establiment, ja sigui per a nombre d'establiments o de places.

Nombre de paso: Normalización de Fila
Tipo de campo: tipus_dada

Campos

#.	Nombre campo	Tipo	campo nuevo
1	n_hotels	n_hotels	quantitat
2	n_campings	n_campings	quantitat
3	n_turisme_rural	n_turisme_rural	quantitat
4	n_total	n_total	quantitat
5	n_hotels_estrelles_or	n_hotels_estrelles_or	quantitat
6	n_hotels_estrelles_argent	n_hotels_estrelles_argent	quantitat
7	n_hotels_total_estrelles	n_hotels_total_estrelles	quantitat
8	n_campings_cat_1	n_campings_cat_1	quantitat
9	n_campings_cat_2	n_campings_cat_2	quantitat
10	n_campings_cat_3	n_campings_cat_3	quantitat
11	n_campings_cat_privat	n_campings_cat_privat	quantitat
12	n_campings_total_cat	n_campings_total_cat	quantitat
13	n_turisme_rural_allotjament_independent	n_turisme_rural_allotjament_independent	quantitat
14	n_turisme_rural_masia	n_turisme_rural_masia	quantitat
15	n_turisme_rural_casa_poble	n_turisme_rural_casa_poble	quantitat
16	n_turisme_rural_total	n_turisme_rural_total	quantitat
17	p_hotels	p_hotels	quantitat
18	p_campings	p_campings	quantitat
19	p_turisme_rural	p_turisme_rural	quantitat
20	p_total	p_total	quantitat
21	p_hotels_estrelles_or	p_hotels_estrelles_or	quantitat
22	p_hotels_estrelles_argent	p_hotels_estrelles_argent	quantitat
23	p_hotels_total_estrelles	p_hotels_total_estrelles	quantitat
24	p_campings_cat_1	p_campings_cat_1	quantitat
25	p_campings_cat_2	p_campings_cat_2	quantitat
26	p_campings_cat_3	p_campings_cat_3	quantitat
27	p_campings_cat_privat	p_campings_cat_privat	quantitat
28	p_campings_total_cat	p_campings_total_cat	quantitat
29	p_turisme_rural_allotjament_independent	p_turisme_rural_allotjament_independent	quantitat
30	p_turisme_rural_masia	p_turisme_rural_masia	quantitat
31	p_turisme_rural_casa_poble	p_turisme_rural_casa_poble	quantitat
32	p_turisme_rural_total	p_turisme_rural_total	quantitat

Per a corregir les dades he usat la funció “Valor Java Script Modificado”, ha estat necessari

conèixer les particularitats de JavaScript i del seu sistema de variables.
Aquestes dades emmagatzemades en taules de MySQL m'han servit per extreure les dades usant funcions PL/SQL, que poden funcionar amb paràmetres. Ha estat necessari aprendre PL/SQL per a MySQL.
Els informes els he creat usant *Pentaho Report Designer*, que m'ha permès usar paràmetres a les funcions PL/SQL. En aquesta eina és a on es veu que funciona tot el treball realitzat. Han de funcionar en conjunt la Base de Dades, les funcions PL/SQL i els paràmetres inclosos a l'informe.

Ho he fet tot amb programes, sense tocar res manualment. Això ha suposat una complicació extra.

He afegit un informe extra que permet visualitzar un mapa de la localització dels equipaments, per a fer-ho he usat *html*, *javascript* i *Pentaho Report Designer.*

3.5 Conclusions

En aquest projecte ens podem adonar del servei que un informàtic amb coneixements de magatzems de dades pot aportar a una empresa per millorar el seu funcionament, la seva rendibilitat i la seva viabilitat. Ja que les dades es poden analitzar per extreure informació útil per a la presa de decisions. Aquesta informació reflectirà tendències i això podrà ser aprofitat per tenir avantatge sobre els competidors.

A la vista de l'anàlisi de dades feta a l' apartat 3.2.4.3 d'anàlisi de les dades, en els processos ETL hi ha una gran feina de transformació de les dades.

En quant a l'aprenentatge al llarg del TFC, s'ha de fer un esforç considerable partint de no tenir coneixements sobre Magatzems de Dades ni del programari relacionat, principalment Pentaho.
S'ha de tenir cautela en el plantejament de funcionalitats perquè el desconeixement del tema fàcilment pot suposar entrebancs insuperables o que facin perdre un temps que excedeixi el normal per a una assignatura de 7,5 crèdits, de fet es va preveure gairebé el doble del temps que normalment s'ha de preveure per cada crèdit.

Per a fer l'anàlisi i disseny convé que es coneguin amb certa profunditat els temes de Magatzems de dades i que s'hagin provat les possibilitats de les eines amb les que s'ha de treballar a la fase d'implementació. Aquestes dues coses requereixen una forta inversió de temps.

3.6 Línies d’evolució futura

Es podria fer una taula comú de municipis però s'hauria de normalitzar el nom dels municipis i obtenir dades dels municipis nous provinents de la taula 'Equipament'.
Es podrien fer Dashboards amb gràfics de les dades.
Quadres de comandament.
Anàlisi OLAP.
Diverses millores que una vegada arribat el punt de desenvolupament en que està el projecte no s'han fet per què el temps no ho ha permès. Per exemple un filtre a l'informe extra sobre localització d'equipaments.
Incorporació de dades que no venien en els arxius aportats per ONdO.

4. Glossari

Business Intelligence: (Intel·ligència empresarial). Es denomina intel·ligència empresarial, intel·ligència de negocis o BI (de l'anglès *business intelligence*) al conjunt d'estratègies i eines enfocades a l'administració i creació de coneixement mitjançant l'anàlisi de dades existents en una organització o empresa.

Magatzem de dades: Sistema informàtic utilitzat com a eina de suport per a la presa de decisions que integra una gran diversitat d'informació procedent de diferents bases de dades i que permet realitzar consultes complexes i de tipus analític sobre aquesta informació.
Font: http://www.termcat.cat

ETL: *Extract, Transform and Load* (Extreure, transformar i carregar en angles, freqüentment abreviat a ETL) és el proces que permet a les organitzacions moure dades des de múltiples fonts, re-formatar-les i netejar-les, i carregar-les en una altra base de dades, *data mart*, o *data warehouse* per a analitzar, o en un altre sistema operacional per a donar suport a un proces de negoci.
Font: http://es.wikipedia.org/wiki/Extract,_transform_and_load

OLAP: *On-Line Analytical Processing* (acrònim en angles de processament analític en línea). És una solució usada en el camp de l'anomenada Intel·ligència empresarial (o Business Intelligence) l'objectiu de la qual es agilitzar la consulta de grans quantitats de dades.
Font: http://es.wikipedia.org/wiki/OLAP

5. Bibliografia

Pla docent de l'assignatura.
Centre de software de Ubuntu.
http://es.wikipedia.org/
Materials sobre Atlas SBI.
http://www.termcat.cat
Materials de les assignatures BD I i BD II de la UOC.
Materials de l'assignatura Magatzems de dades i models multidimensionals de la UOC.
Centre de Software de Ubuntu.
www.pentaho.com

6. Annexos

6.1 Enunciat

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

Enunciat

El nombre d'allotjaments turístics a Catalunya ha seguit creixent durant l'any 2012. Davant d'aquesta situació, l'Observatori Nacional d'Ocupació (ONdO) vol 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.

A l'hora de recollir les dades necessàries, ONdO ha sol·licitat a l'IDESCAT informació sobre població i equipaments (biblioteques, teatres, parcs...), i a la Federació Catalana d'Allotjaments Turístics la informació sobre establiments (hotels, càmpings i turisme rural), les places ofertes i les característiques dels establiments catalans.

L'ONdO ha decidit encarregar-nos, com a consultora externa independent, la creació d'un magatzem de dades per obtenir, com a mínim informació relativa a:

- ✓ 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

Tota aquesta informació es podrà consultar de forma agregada, per comarca/província, tipus d'establiment i categoria. La temporalitat de les dades serà a nivell d'any.

També haurem de proporcionar un conjunt prefixat d'informes on es mostri la informació sol·licitada i qualsevol altre que creguem que pugui ser útil per a l'ONdO

ONdO ens proporcionarà tota la informació rebuda en els següents fitxers:

- ✓ *Establiments XXXX*: Un arxiu per cada any amb: nombre d'establiments, tipologia, places ofertes i classificació
- ✓ *Equipaments*. Arxiu amb la relació d'equipaments públics a data 31/12/2012.
- ✓ *Poblacio*: Un arxiu amb els habitants per any i la superfície de cada municipi.

Ens adverteixen que degut a que la informació s'ha extret de diferents sistemes, podem tenir formats csv diferents (separat per coma, tabulador i punt i coma).

Finalment ens demanen que calculem el nombre d'habitants com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent.

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 l'ONdO en format text. 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 següent tasques:

- ⇨ 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 maquinari i programari

Es treballarà sobre una màquina virtual de VirtualBox proporcionada per la UOC.