CONSTRUCCIÓ I EXPLOTACIÓ D'UN MAGATZEM DE DADES PER A L'ANÀLISI D'INFRAESTRUCTURES TURÍSTIQUES I PERNOCTACIONS

Vadym Andriyevskyy
Enginyeria Informàtica en Tecnologies de la Informació

Carles Llorach Rius

16 de juny del 2015

FITXA DEL TREBALL FINAL

Títol del treball:	Construcció i explotació d'un magatzem de dades per a l'anàlisi d'infraestructures turístiques i pernoctacions
Nom de l'autor:	Vadym Andriyevskyy
Nom del consultor:	Carles Llorach Rius
Data de lliurament (mm/aaaa):	06/2015
Àrea del Treball Final:	Magatzems de dades
Titulació:	Enginyeria Informàtica en Tecnologies de la Informació

Resum del Treball:

Estem en un moment del mercat globalitzat, dinàmic i canviant. L'anomenada "societat d'informació" ens sobrecarrega de dades. En el sector empresarial el coneixement és vital per seguir endavant o tenir avantatge competitiu. Però totes les dades són inútils si no podem extreure d'elles un coneixement fiable. Per ajudar en la presa de decisions estratègiques cada vegada pren més protagonisme la disciplina de Business Intelligence.

Aquest document correspon a la memòria del treball final de grau enfocat al desenvolupament d'un projecte de Datawarehouse per una empresa del sector turístic que vol fer un estudi a Catalunya per determinar la continuïtat de la seva xarxa d'establiments, coneixent millor els criteris decisoris dels seus clients i els motius de l'evolució negativa de pernoctacions dels darrers anys.

El treball es centra en la creació i exploració d'un Datawarehouse, incloent la seva arquitectura bàsica – processos de tractament i entrada de dades, el seu emmagatzemament i extracció del coneixement.

En aquesta memòria s'exposen totes les fases vitals seguides en el desenvolupament del projecte, des de la planificació fins a la implementació, passant per anàlisi i disseny, tot enfocat segons els requeriments proporcionats pel client final. El resultat final és una solució integrada i implementada en una màquina virtual a punt per la seva explotació.

Abstract:

Nowadays, we are facing a moment of global market, dynamic and continuously changing. The so called “information society” is flooding us with data. In business sector, the knowledge is vital to move forward or get some competitive edge. However, all the data is useless if we are not able to extract from it the reliable knowledge. The discipline of Business Intelligence, which helps to take strategic decisions, is becoming more and more prominence.

This document corresponds to a dissertation focused on the development of a Datawarehouse for a touristic company which wants to perform a study in Catalonia to determine the continuity of its network of establishments, improving their knowledge regarding the criteria in decision making of their clients and the reasons for the negative development of overnight stays in the last years.

The whole work is focused on the creation and exploration of a Datawarehouse, including its basic architecture – data treatment and load, its storing and extraction of knowledge.

In this dissertation are exposed all the vital stages in the development of the project, from planning, followed by analysis and design, to implementation, altogether focused on the requirements of the final client. The final result is an integrated solution implemented on a virtual machine and ready for operation.

Paraules clau (entre 4 i 8):

Data Warehouse, Magatzem de Dades, Business Intelligence, ETL, OLAP, ROLAP, SQL, Pentaho.

ÍNDEX

LLISTA DE FIGURES

1. INTRODUCCIÓ

1.1. CONTEXT I JUSTIFICACIÓ DEL TREBALL

Actualment una organització gira al voltant de la informació. Es reben dades per tot arreu, es processen i es fabriquen de noves. Sovint aquestes dades es distribueixen de forma ineficient entre els diferents sistemes dels que es disposa. La gestió es torna tediosa, amb problemes de funcionament general, donant lloc a improvisacions. La integració de diferents sistemes d'informació s'ha convertit en un repte constant.

D'altra banda la competència és feroç. Per aconseguir un avantatge competitiu no només és imprescindible saber tractar correctament totes les dades disponibles, sinó que també cal fer-ho de manera ràpida i eficient, per així tenir totes les bases per prendre les decisions adequades en el mínim temps possible.

D'aquesta manera el Data Warehouse (DW) es converteix en un element essencial per resoldre el problema de treure les dades interessants dels sistemes transaccionals de manera ràpida i eficient i convertir-los en informació processable. Separat dels sistemes operacionals es dissenya per a donar suport a les decisions, treure informes analítics, consultes ad-hoc i mineria de dades. En definitiva, el DW és un fonament bàsic per l'èxit de la intel·ligència de negoci (BI) empresarial.

S'ha plantejat com a projecte de final de grau la construcció d'un DW amb la intenció d'aprofundir en aquest nou camp dins del marc de les bases de dades. Tot el treball està projectat des d'una perspectiva d'estudiant que a partir d'uns coneixement bàsics, la recerca d'informació i el treball diari es proposa construir un magatzem de dades similar al que es pot trobar en el món laboral.

1.2. OBJECTIUS DEL TREBALL

1.2.1. GENERALS

L'objectiu principal és desenvolupar un projecte que permeti ampliar els coneixements ja existents de les assignatures d'àrea de les bases de dades, aprofundir en les bases de dades que donen suport a la presa de decisions en les organitzacions, al mateix temps que aplicar en un entorn pràctic altres habilitats adquirides al llarg dels estudis de grau.

Més concretament es tracta 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.

Objectius a nivell personal:

- Posar en pràctica la correcta gestió de projectes, tot seguint les tècniques estudiades en l'assignatura de Gestió de projectes.
- Enfrontar-se a una situació similar a la que es pugui trobar en el mon laboral.
- Aprendre a gestionar el temps, els recursos i els cicles d'un projecte tecnològic.
- Raonar profundament sobre els riscos, aplicant les mesures de contenció necessàries.

Objectius a nivell de projecte:

- Comprendre què és un magatzem de dades i quins avantatges aporta enfront les bases de dades operacionals.
- Saber fer un disseny multidimensional a partir d'un conjunt de requisits que descriuen una problemàtica donada.
- Conèixer els problemes de la integració, transformació i càrrega de dades i saber resoldre'ls per a crear un magatzem de dades a partir de múltiples fonts.
- Saber crear un magatzem de dades partint d'un disseny multidimensional i implementar-lo fent servir la tecnologia que es consideri més adequada.
- Conèixer aplicacions i eines per a una òptima explotació del magatzem de dades i utilitzar alguna d'elles per a donar resposta al problema plantejat.

1.2.2. TECNOLÒGICS

El projecte gira entorn el disseny i la implementació d'un magatzem de dades, que és l'objectiu principal del projecte, juntament amb els components necessaris per la seva explotació, entre ells els més importants que cobreixen les funcionalitats:

- Càrrega de dades.
- Exploració i anàlisi de dades.
- Visualització de dades.

S'ha fixat com a prioritat que el màxim de components situats abans, durant i després del DW siguin de codi lliure, amb un ampli suport de la comunitat.

1.3. ENFOCAMENT I MÈTODE SEGUIT

El projecte s'ha desenvolupat seguint majoritàriament una metodologia tradicional de gestió de projectes, amb alguns petits trets de les metodologies adaptatives que han permès una certa flexibilitat. No obstant aquestes

desviacions, el guió principal de la metodologia tradicional no ha variat, ja que el projecte s'ha hagut d'alinear amb les entregues de les PAC, fixades des del principi.

D'aquesta manera s'han seguit les següents fases del projecte:

1. **Inicialització**. En aquest apartat s'ha introduït a la problemàtica plantejada, fent un anàlisi preliminar de requisits amb la possible arquitectura a implementar.
2. **Planificació**. Aquí s'han detallat els recursos necessaris, s'ha distribuït el temps disponible i s'han identificat els possibles riscos juntament amb els seus mètodes de contingència.
3. **Anàlisi detallat i disseny**. En aquest apartat s'ha entrat a fons en tots els requeriments necessaris, que han permès entendre correctament tots els detalls del projecte final. Seguidament s'ha elaborat el disseny de les parts vitals del projecte segons l'arquitectura proposada a l'inici.
4. **Execució**. En aquesta fase s'han implementat, en un entorn real, els dissenys de la fase anterior, incloent-hi tots els elements necessaris per la seva correcta execució. També s'han elaborat els informes demanats.
5. **Monitoratge i control**. Aquesta fase ha consistit en un procés permanent i paral·lel a tot el projecte. Tots els aspectes continguts en la planificació inicial s'han avaluat i reajustat en cas de necessitat.
6. **Tancament**. En aquest últim apartat s'han fet tots els treballs necessaris per poder fer l'entrega del producte final.

Durant el desenvolupament del projecte, els apartats considerats com a vitals i els que comporten carrega de feina més gran, van rebre una atenció especial. Són els que es llisten a continuació:

- Disseny multidimensional.
- Extracció, transformació i càrrega de dades.
- Generació d'informes.

1.4. PLANIFICACIÓ DEL TREBALL

1.4.1. RECURSOS

1.4.1.1. HW/SW

El producte final està desenvolupat sobre un entorn virtual hostejat a Amazon, proporcionat per la UOC. A efectes pràctics les característiques d'entorn són:

- Processador: Intel Xeon E5-2680 a 2,80GHz
- Memòria RAM: 3,75GB
- Espai lliure: 9GB

L'entorn té instal·lats els següents components:

- Sistema operatiu: Windows Server 2012 64bits

- DBMS encarregat d'allotjar el DW: MySQL Server
- Suite BI per l'elaboració d'informes: Pentaho Business Analytics

El treball diari, probes i documentació s'han fet sobre una màquina física amb les següents característiques:

- Processador: Intel Core i5-2430M a 2,40GHz
- Memòria RAM: 8GB
- Sistema operatiu: Windows 7 SP1 64bits

1.4.1.2. Temps

El meu temps s'ha dividit entre la feina, el TFG de 12 crèdits i una altra assignatura de 6 crèdits. L'horari oficial dedicat a la UOC és de dimarts a divendres de 16:00 a 21:00 i un dia complet de fins a 10 hores el dissabte o el diumenge. D'aquesta manera s'ha disposat d'unes 30 hores setmanals. Fent la proporció de crèdits al TFG, es correspon a unes 20 hores cada setmana.

Seguint el calendari establert la distribució de temps ha estat la següent:

RECURS	Quantitat
Hores totals	**300**
Pla de treball	30
Anàlisi i disseny	120
Implementació	120
Memòria i presentació virtual	30

Cal mencionar que el consultor també ha dedicat una part del seu temps al seguiment i supervisió del projecte. Aquestes hores no estan recollides dins del càlcul. També és important destacar que el meu horari va ser flexible i s'ha adaptat a cada moment del treball. Així, per exemple, s'han utilitzat moltes hores per les mesures de contingència de riscos.

1.4.2. TASQUES I FITES

Paquet de treball	Pla de treball i anàlisi preliminar de requeriments		
Data inici	27/02/2015	**Data fi**	11/03/2015
Recursos previstos	30 hores		
Objectius			
- Planificar i estructurar el projecte. - Elaborar un anàlisi preliminar general. - Analitzar les fonts de dades.			
Descripció de la tasca			
Es descriu de forma clara el problema que pretén resoldre el projecte, el treball concret que es portarà a terme i la seva descomposició en tasques i fites temporals. També s'analitzen les fonts de dades proporcionades.			
Entregables			**Termini**
PAC 1			11/03/2015

Paquet de treball	Anàlisi de requeriments i disseny conceptual i tècnic		
Data inici	12/03/2015	**Data fi**	15/04/2015
Recursos previstos	120 hores		
Objectius			
- Anàlisi detallat de requeriments - Disseny del model dimensional - Disseny dels procediments d'extracció de dades			
Descripció de la tasca			
Es fa un anàlisi detallat de requeriments basat en l'anàlisi preliminar realitzat. També es realitza el disseny del model dimensional que donarà suport a les necessitats dels usuaris. Per últim es defineixen els procediments d'extracció de dades a alt nivell.			
Entregables			**Termini**
PAC 2			15/04/2015

Paquet de treball	Implementació		
Data inici	16/04/2015	**Data fi**	28/05/2015
Recursos previstos	120 hores		
Objectius			
- Construcció del magatzem de dades - Càrrega de dades - Implementació d'eina OLAP - Construcció d'informes - Anàlisi de la informació			
Descripció de la tasca			
Es construeix el magatzem de dades amb tots els seus components seguint els dissenys marcats i es fa la càrrega de dades. També s'implementen les eines d'explotació, informes i anàlisi d'informació.			
Entregables			**Termini**
PAC 3			28/05/2015

Paquet de treball	Lliurament final		
Data inici	29/05/2015	**Data fi**	16/06/2015
Recursos previstos	30 hores		
Objectius			
- Producte a punt per entregar - Memòria completada - Presentació virtual			
Descripció de la tasca			
Es poleix el producte final i es prepara la seva entrega. Es confecciona la memòria del treball seguint les indicacions del pla docent. Per últim, es prepara una presentació virtual del treball valorant els objectius aconseguits i es treuen les conclusions pertinents.			
Entregables			**Termini**
PAC4	Producte		16/06/2015
	Memòria		16/06/2015
	Presentació virtual		16/06/2015

1.4.3. *DIAGRAMA DE GANTT*

Id	Nombre de tarea	Comienzo	Fin
1	**TFG - Data warehouse**	**vie 27/02/15**	**mar 16/06/15**
2	**PAC1 - Pla de treball**	**vie 27/02/15**	**mié 11/03/15**
3	Estudi de conceptes	vie 27/02/15	vie 06/03/15
4	Estudi del cas	vie 27/02/15	mar 03/03/15
5	**Elaboració del pla**	**mié 04/03/15**	**mar 10/03/15**
6	Definició d'objectius	mié 04/03/15	mié 04/03/15
7	Estudi tècnic	jue 05/03/15	vie 06/03/15
8	Repartiment del temps i valoració	lun 09/03/15	mar 10/03/15
9	Correcció i lliurament	mié 11/03/15	mié 11/03/15
10	**PAC2 - Anàlisi i disseny**	**jue 12/03/15**	**mié 15/04/15**
11	Estudi del disseny de DW	jue 12/03/15	dom 15/03/15
12	Anàlisi de requeriments	lun 16/03/15	mié 18/03/15
13	Estudi del model dimensional	jue 19/03/15	mié 25/03/15
14	Disseny de la base de dades	lun 23/03/15	vie 27/03/15
15	Estudi i selecció d'eines ETL	sáb 28/03/15	jue 02/04/15
16	Disseny ETL	mar 31/03/15	vie 03/04/15
17	Estudi i selecció d'eines OLAP	sáb 04/04/15	jue 09/04/15
18	Disseny d'OLAP	mar 07/04/15	sáb 11/04/15
19	Disseny d'informes	dom 12/04/15	lun 13/04/15
20	Correcció i lliurament	lun 13/04/15	mié 15/04/15
21	**PAC3 - Implementació**	**jue 16/04/15**	**jue 28/05/15**
22	Configuració d'entorn	jue 16/04/15	vie 17/04/15
23	Creació i test de la base de dades	vie 17/04/15	lun 27/04/15
24	Configuració, càrrega i test d'ETL	mar 28/04/15	mié 06/05/15
25	Configuració i test d'OLAP	jue 07/05/15	dom 17/05/15
26	Creació d'informes	lun 18/05/15	vie 22/05/15
27	Anàlisi d'informació	vie 22/05/15	dom 24/05/15
28	Correcció i lliurament	lun 25/05/15	jue 28/05/15
29	**PAC4 - Lliurament final i defensa**	**vie 29/05/15**	**mar 16/06/15**
30	Preparació del producte a lliurar	vie 29/05/15	dom 31/05/15
31	Memòria de treball	lun 01/06/15	dom 07/06/15
32	Presentació virtual	lun 08/06/15	dom 14/06/15
33	Correcció i lliurament	sáb 13/06/15	mar 16/06/15

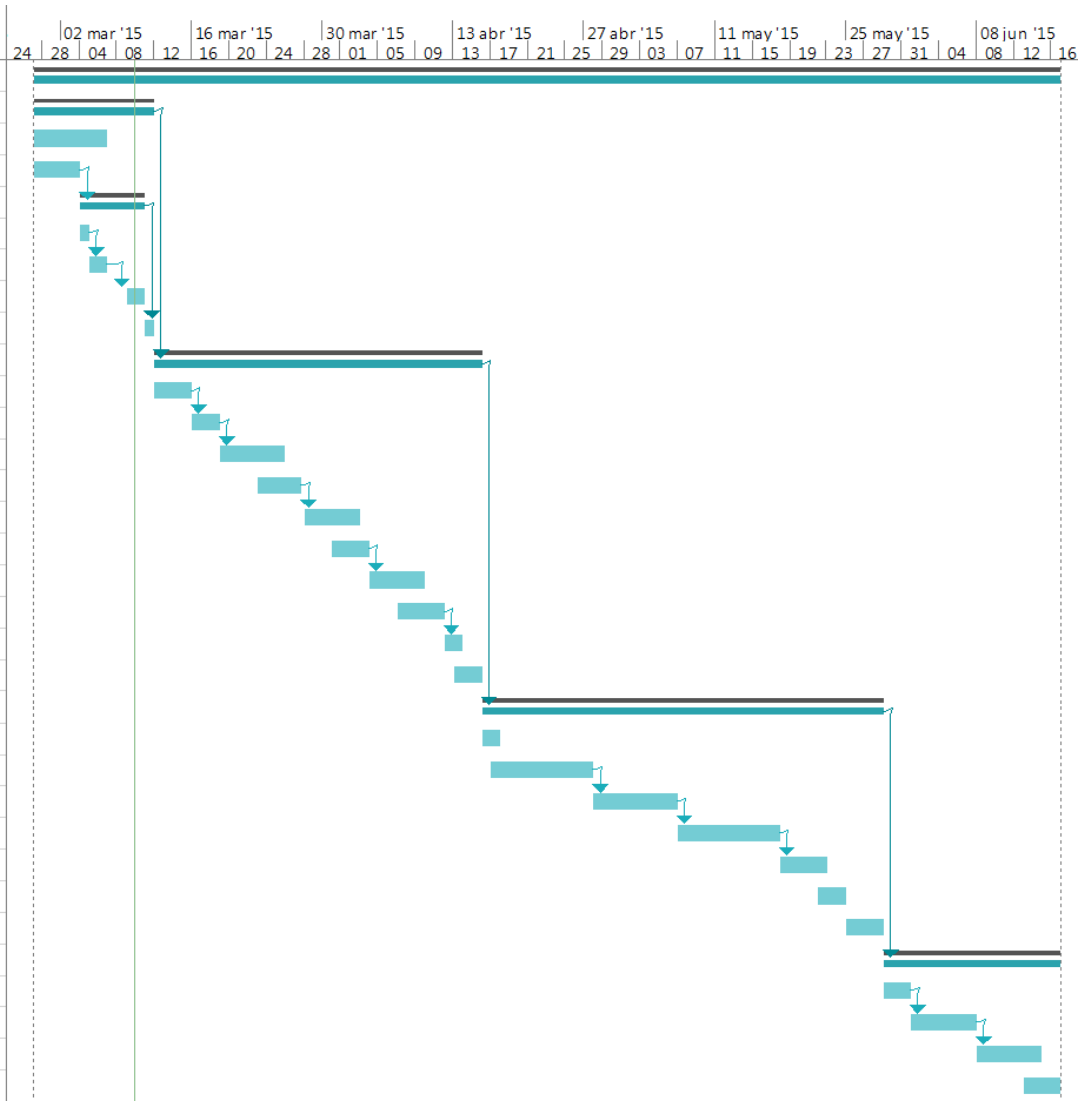


Figura 1. Diagrama de Gantt

1.4.4. ANÀLISI DE RISCOS

A continuació es detallen els diferents riscos previsibles tant externs com interns que poden sorgir durant el desenvolupament del TFG, juntament amb el pla de contingència proposat.

RISCOS EXTERNS	
Incidència	**Pla de contingència**
Reducció de la disponibilitat horària	La disponibilitat horària pot ser afectada per diversos raons laborals, educatius o personals. La solució passaria per dedicar-hi més hores el cap de setmana i en cas de necessitat es demanarien vacances a la feina per poder dedicar més temps al projecte.
Malaltia	En cas d'una malaltia lleu el treball continuaria endavant. En cas de ser possible es posposarien feines més intensives i es farien revisions i investigacions. En cas d'una malaltia més greu es procediria a aplicar el pla de contingència del punt anterior, si no és possible es comunicaria directament al consultor i conjuntament es decidiria la continuïtat del projecte.
Avaria informàtica	Es configura l'entorn de treball amb còpies en el núvol immediates. Es procedeix immediatament a fer les reparacions necessàries. També es disposa d'un equip de substitució.

RISCOS INTERNS	
Incidència	**Pla de contingència**
Contratemps en entrega de les PAC	Com a primera mesura preventiva es posa com a objectiu tenir totes les PAC fetes com a molt tard un dia abans de la data d'entrega. El dia restant es dedicarà al repàs i millora. En cas d'un gran contratemps es dedicarien més hores el cap de setmana o directament es demanarien vacances a la feina.
Estancament en un punt del projecte	En cas d'estancament es procedirà a identificar el punt problemàtic, dedicar-hi més temps buscant solució, exposar-ho al consultor. En cas d'un punt mort es buscaran alternatives.

1.4.5. PLANIFICACIÓ FINAL

L'execució final ha estat majoritàriament fidel a la inicial, però amb algunes incidències importants.

En primer lloc, per raons familiars i laborals pràcticament no s'ha pogut disposar dels caps de setmana. En segon lloc, he passat dues malalties que m'han impedir dedicar temps a la feina durant 3 setmanes. Per últim, hi ha hagut un desbordament de la càrrega de treball prevista inicialment en l'altra assignatura de la UOC que compagino amb TFG.

Tot això m'ha obligat a modificar la planificació inicial, traient dels requisits tota la part de cubs OLAP, ja que aquesta no era obligatòria. Per aconseguit complir amb les dates de les PAC i les tasques a realitzar s'han augmentat les hores de treball diàries, afegint també els dilluns i agafant dues setmanes de vacances a la feina distribuïdes entre les fases de disseny i implementació.

En definitiva, s'han aplicat les mesures de contingència previstes per fer front a les incidències i modificar el temps disponible, i complir així amb les dates d'entrega marcades.

1.5. BREU SUMARI DE PRODUCTES OBTINGUTS

Els productes obtinguts són de dos tipus. Els primers són els relacionats amb l'avaluació contínua.

1. PAC1 – Pla de treball.
2. PAC2 – Anàlisi i disseny.
3. PAC3 – Implementació, accés a les dades i informes creats.
4. PAC4 – Memòria, autoinforme i presentació virtual.

El segon tipus són els productes resultants, que inclouen:

1. Base de dades en MySQL. Amb l'estructura i els usuaris creats i les dades carregades.
2. Processos ETL en Pentaho Spoon. Es tracta de tots els processos necessaris que extreuen les dades dels fitxers disponibles, les netegen, transformen i carreguen a la base de dades.
3. Informes. S'entreguen tant els dissenys dels informes, com aquests carregats dins de la suite Pentaho BI i accessibles des del navegador als usuaris finals.

1.6. Breu descripció dels altres capítols de la memòria

La resta de capítols de la memòria estan estructurats segons l'execució del projecte.

Primer s'exposa l'anàlisi detallat de tot el projecte. Aquest apartat s'introdueix amb la descripció tècnica, que explica des de la situació general fins la solució proposada. Es prossegueix amb un anàlisi exhaustiu de requeriments que s'han demanat per complir els objectius del projecte. Finalment es desglossen les fonts de dades entregades pel Grup Líder de Turisme Familiar (GLTF).

Una vegada completat l'anàlisi, es passa al disseny de la solució. Primer s'exposa el pilar de tot el projecte – el disseny multidimensional de la base de dades, i després els processos ETL, que degut a la seva complexitat requereixen de molta planificació abans de ser implementats.

El següent capítol s'ha dedicat a la implementació de la solució dissenyada. Comença explicant totes les tasques relacionades amb la base de dades, prosseguint amb els processos ETL i informes de cara a l'usuari, per finalitzar amb aquelles tasques necessàries per llençar el servidor a la producció.

Es conclou amb la projecció de les millores i les futures línies de treball a seguir, juntament amb les conclusions finals.

2. ANÀLISI

2.1. DESCRIPCIÓ TÈCNICA DEL PROJECTE

2.1.1. PRECEDENTS

Se'ns posa davant d'una situació d'una empresa fictícia, amb una problemàtica concreta. Cal actuar com una consultoria externa independent que ha de crear els productes amb l'objectiu d'ajudar a entendre més bé la situació actual a partir d'una sèrie d'indicadors i informes prefixats.

És important destacar que al ser una empresa fictícia, la consultoria (és a dir, l'estudiant), no té un tracte directe amb el client, no hi ha reunions i aprovacions per part d'aquest. Per tant tot el treball es basa en els requeriments descrits en l'enunciat, i el consultor és el que ha controlat i validat el rumb del projecte.

2.1.2. ESTAT ACTUAL I PROBLEMÀTICA

Estem en una situació de crisi i un dels sectors que la pateix és el de turisme. El Grup Líder en Turisme Familiar (GLTF) ha detectat una evolució negativa de pernoctacions dels darrers anys. El problema és que fins ara no s'ha observat una correlació clara entre els elements que caracteritzen un municipi (equipaments, infraestructures disponibles, percepció de seguretat...) i les pernoctacions que s'hi esdevenen.

Degut a aquesta problemàtica s'ha decidit fer un estudi delimitat a la regió geogràfica de Catalunya, que ha de permetre visualitzar els criteris decisoris dels seus clients i els motius de l'evolució negativa de les pernoctacions. En definitiva, per tal de determinar les accions que cal dur a terme necessita tenir una visió més detallada de la realitat.

2.1.3. SOLUCIÓ PROPOSADA

Per tal de satisfer les necessitats del GLTF es crea un magatzem de dades a través del qual es pugui obtenir informació completa, a fi de realitzar les decisions estratègiques necessàries.

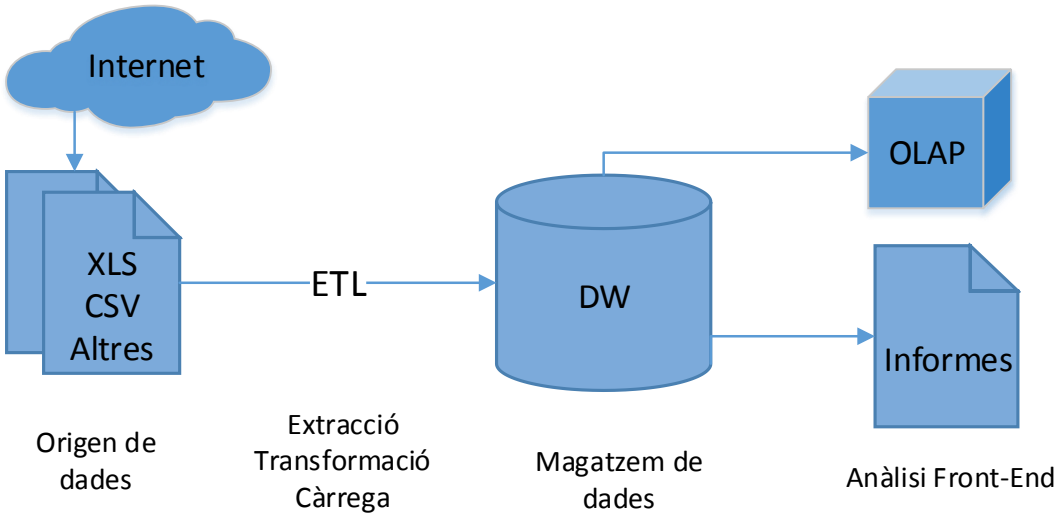


Figura 2. Esquema general del projecte

Origen de dades

Es disposa de la mateixa informació que té el GLTF en els seus fitxers de text. No obstant, per realitzar l'anàlisi demanat aquesta informació no és suficient, per tant s'han incorporat algunes dades de les fonts d'Internet.

Extracció, Transformació i Càrrega

Les dades provinents del GLTF de les que es disposa originàriament s'han extret de diferents sistemes i no estan normalitzades, ni tant sols estan en els mateixos formats de fitxer. S'ha implementat un sistema ETL que fa una neteja de dades extraient només les importants, les transforma al format destí i les insereix en el DW, tot i que alguns fitxers no tenen una estructura lògica clara i per això s'han hagut de tornar a extreure de les seves fonts originals.

Magatzem de dades

Està implementat sobre un sistema DBMS relacional de codi obert, seguint un disseny multidimensional. No es contempla cap implementació de repositori de metadades, ni divisió del DW en Data marts.

Anàlisi Front-End

És possible tenir diferents tipus de sortides del sistema. En el projecte s'han implementat els informes com a sortida principal. Aquests s'encarreguen de mostrar la informació sol·licitada expressament pel client, juntament amb tota aquella que sigui considerada útil.

Com a alternativa també es podrien implementar els cubs OLAP que permeten extreure tot el potencial analític amagat en la típica estructura dimensional del DW. Aquesta línia de treball no està implementada, però sí projectada com a futura línia de desenvolupament.

2.2. ANÀLISI DE REQUERIMENTS

A continuació es repassen els resultats mínims que s'espera obtenir. A partir d'aquests desglossarem la informació que es requereix i els processos que cal seguir per obtenir-los.

Un apunt important a destacar és que l'estudi es farà a partir de les dades dels anys 2011, 2012 i 2013, sent el 2013 l'últim any del que podem disposar de dades definitives.

2.2.1. Rànquing de municipis per categoria d'equipaments.

Aquí es toca el primer tema general – els municipis. Cada municipi es defineix per un nom i un codi postal, que és la única informació del municipi que necessitem en aquest indicador.

D'altra banda se'ns demana que ho relacionem amb els equipaments municipals. Segons la definició[1], aquests són un conjunt d'espais i edificacions, majoritàriament d'ús públic, en els que es realitzen activitats complementàries a l'habitatge i feina, o bé, en els que es proporciona a la població serveis de benestar social i suport per les activitats econòmiques. Es classifiquen en funció d'activitats o serveis que proporcionen. D'aquesta manera cada equipament pertany a una o diverses categories determinades.

És important destacar que els equipaments s'ordenen per jerarquia. Com a mínim sempre hi ha dos nivells jeràrquics – un que indica el camp d'actuació i un altre que indica l'especificació. Evidentment es podrien desglossar els nivells fins l'últim detall, però seria massa costós, ja que no es disposa d'una jerarquia clara, ni de dades contundents. Per aquest motiu s'han considerat només els dos primers nivells.

Així a nosaltres ens interessa saber el nom de la categoria de cada equipament i el sumatori del número d'equipaments de cada categoria, per cada municipi. L'indicador a més demana fer un rànquing de municipis (millors i pitjors) per cada categoria.

2.2.2. Màxim i mínim d'efectius policials per tipologia d'establiment i municipi.

En aquest indicador se'ns demana relacionar els efectius policials amb els diferents tipus d'establiments i presentar-ho de manera agregada a nivell de municipi, comarca, marca turística o comunitat autònoma.

[1] Secretaría de Asentamientos Humanos y Obras Públicas, Glosario de Términos sobre Asentamientos Humanos, México, 1978

Tenim tres tipus d'establiments turístics: hotels, càmpings i allotjaments rurals. La categoria hotels engloba tots els hotels d'una fins a cinc estrelles, hostals i pensions. La categoria càmpings engloba tots els establiments de càmping de luxe, de primera, segona i tercera categoria. Per últim, la categoria allotjaments rurals engloba les diferents tipus de cases de poble, masies o masoveries.

L'objectiu és fer un promig d'efectius policials per cada tipus establiment, però per ser més precisos també caldria disposar d'un promig per cada plaça d'establiment turístic. Els resultats s'hauran de poder ordenar en ordre ascendent i descendent segons el número d'efectius.

Les dades que es necessiten són els efectius policials que hi ha en cada municipi durant els anys d'estudi, la suma d'establiments turístics de cada tipus i les places amb les que compten cada tipus d'establiments. Per obtenir els indicadors demanats caldrà sumar els policies de cada municipi i fer un promig contra el número de places de cada tipus d'establiment. Els resultats s'hauran de poder ordenar en odre ascendent i descendent. A més, s'hauran de poder fer consultes agregades generalitzades o especificant el lloc de consulta.

2.2.3. Rati de policies locals per habitant.

Bàsicament necessitarem tenir agrupats els policies locals per municipis i fer un promig respecte el cens d'habitants de cada municipi. Caldrà tenir en compte que el cens varia d'un any a l'altre, igual que els policies locals d'un municipi. També cal tenir en compte que si més endavant volem generalitzar l'estudi per comarques, marques turístiques, etc., caldrà recalcular el rati en cada generalització.

2.2.4. Distribució mensual i estacionalitat de les pernoctacions.

El terme pernoctar[2] es defineix com passar la nit en un determinat lloc que no és l'habitatge habitual. En el nostre cas utilitzem les pernoctacions per quantificar les persones que passen la nit en un municipi, però no en són residents habituals.

Primer de tot ens demanen distribuir les pernoctacions mensualment. Com que no s'especifica el nivell geogràfic d'aquesta distribució caldrà fer-la a nivell de la marca turística, ja que aquest és el nivell de dades amb el que treballa l'INE[3].

També ens demanen l'estacionalitat de les pernoctacions. El terme estacionalitat[4] o variació estacional es defineix com la variació periòdica i predictible d'una sèrie temporal en un període que es comprèn dins d'un any. Per detectar l'estacionalitat caldran eines estadístiques prou potents que

[2] Real Academia Española. Diccionario Usual.
[3] Instituto Nacional de Estadística.
[4] Lexicoon.org – diccionari i traductor.

permetin trobar models matemàtics a partir de la sèrie temporal disponible (de tres anys), separant la tendència, l'estacionalitat i altres fluctuacions irregulars. D'aquesta manera no només es sabrà l'estacionalitat de les pernoctacions, sinó que es podrà predir amb poc marge d'error l'evolució de les nostres pernoctacions al llarg dels pròxims anys.

El procediment descrit en el paràgraf anterior és molt costos d'elaborar i segurament es necessitaria d'una suite de BI. En el nostre cas no disposem dels recursos (temps i diners) necessaris per dur a terme aquesta implementació. Per això l'alternativa és alimentar el sistema d'estacionalitats predefinides. D'aquesta manera es revisarà el calendari dels anys 2011, 2012 i 2013 buscant les estacionalitats que hi ha hagut. Cada període estacional (assignat a mesos concrets) tindrà assignat un pes, així es podran categoritzar les diferents estacions, i si algunes estacions es solapen només caldrà fer una suma de pesos.

2.2.5. % de pernoctacions d'un municipi sobre el total.

Una vegada definides i distribuïdes correctament les pernoctacions, el càlcul del seu percentatge és un procediment simple, dividint les pernoctacions d'una zona geogràfica (mínima del municipi) sobre les pernoctacions totals i multiplicant per cent.

El problema amb el que ens trobem en aquest apartat és la inexistència de dades de pernoctacions segons el municipi, ja que l'INE treballa a nivell de marca turística. Per tant, s'haurà de buscar una estratègia vàlida per distribuir les dades en municipis de forma manual. Se'n suggereixen dues:

- Distribuir les pernoctacions segons el cens municipal.
- Distribuir les pernoctacions segons el número de places hoteleres de cada municipi.

El cens municipal no té una relació directa amb el turisme, però el número de places hoteleres sí. Per això es seguirà la segona estratègia, tenint present en tot moment que les dades de pernoctacions segons el municipi i la comarca no seran exactes, sinó només una aproximació.

Cal tenir en compte que es treballa amb sèries temporals, així que la temporalitat és vital. El percentatge demanat en aquest indicador es pot calcular comprenent com a franja de pernoctacions mensual, bimensual, trimestral, etc. Les pernoctacions variaran segons el sumatori de diferents mensualitats i les operacions posteriors que es facin sobre aquestes dades han de ser coherents amb aquestes temporalitats.

2.2.6. “Top ten” de municipis per franja de pernoctacions.

Aquest indicador és molt similar a l’anterior, només que en lloc de treballar amb percentatges, treballarem amb valors absoluts. Una vegada definida la franja temporal i la zona (municipi) caldrà fer un rànquing ascendent i descendent segons el número de pernoctacions.

2.2.7. Total de pernoctacions estimades per municipi.

El pronòstic de les sèries temporals és molt complex i requereix de la utilització d’eines d’estadística avançades. En el nostre model simplificarem molt el càlcul, deixant de banda la tendència i les fluctuacions irregulars, només aproximant un valor estimat pel municipi.

Quan pensem en la base a partir de la qual hem de calcular les pernoctacions, podríem fàcilment caure en la trampa d’utilitzar només les dades de l’últim any (del mes i municipi que ens interessa). Certament aquestes són molt importants, però poden haver estat alterades per una situació inusual.

Per exemple, el preu del petroli influeix directament en el valor del ruble rus. La baixada del preu del petroli dels últims mesos l’ha devaluat en comparació amb les altres monedes estrangeres, fet que ha provocat una baixada en el turisme rus que previsiblement tornarà als seus números habituals quan el preu del petroli, i de la moneda russa, es recuperi. Per això, si féssim l’estimació d’arribada de turistes russos només a partir de l’últim any, segurament el marge d’error seria molt gran.

Per aquest motiu agafarem com a base del càlcul el promig de pernoctacions dels últims 3 anys, donant més valor als anys més recents.

$$\text{Base} = \text{pernoctacions del } 2011 * 0{,}2 + 2012 * 0{,}3 + 2013 * 0{,}5$$

Com que no utilitzarem estimacions més avançades, la base representarà les pernoctacions estimades.

2.2.8. Promig de viatgers per tipus d’establiment i comarca.

Entenem com a viatger una persona que realitza una o diverses pernoctacions. Per tant, no podem calcular l’indicador que ens demanen a partir de les pernoctacions i no disposem de cap estudi que ens pugui indicar la relació entre pernoctacions per viatger per cada mes donat. Però sí que disposem de números de viatgers entrats cada mes a una zona turística determinada. A partir d’aquí es pot treure el promig de viatgers per comarca. El càlcul es farà seguint la mateixa estratègia que les pernoctacions.

En el cas del promig de viatgers per tipus d’establiment, caldria calcular el promig de viatgers de zones turístiques sobre els diferents tipus d’establiments,

tenint en compte que cada categoria d'establiments té un número diferent de places disponibles.

2.2.9. % d'ocupació per marca turística.

En aquest indicador estem en una situació similar a l'anterior (viatgers-pernoctacions). Per calcular aquest indicador caldrà dividir el número de viatgers que arriben a una marca turística en un mes determinat pel número de places disponibles.

2.2.10. Categorització de municipis A/B/C.

L'objectiu d'aquesta categorització és conèixer els municipis més importants pel que fa a la quantitat de pernoctacions. Això ens permetrà veure quins són els municipis que reben el major nombre de turistes i per tant als que s'haurà de dedicar més atenció, quins són els que tenen potencial de ser importants de cara al negoci i quins altres són residuals i per tant no aporten valor real al negoci.

Segons el principi de Pareto, classificarem els municipis segons les seves pernoctacions en tres segments:

A. El 20% dels municipis que representen el 80% de les pernoctacions. Són els imprescindibles pel negoci, la seva base.
B. El 30% dels municipis que representen el 15% de les pernoctacions. Són els candidats a convertir-se en els importants.
C. El 50% dels municipis que representen el 5% de les pernoctacions restants. Són molts municipis que tenen molt poc moviment turístic i l'atenció als quals ha de ser molt reduïda.

El procediment de categorització es farà ordenant els municipis en ordre descendent i seleccionant els primers X% de municipis que pertanyen a cada categoria.

És important remarcar que la categorització es realitzarà a partir d'un període seleccionat. D'aquesta manera pot ser que per exemple, els municipis líders anuals no siguin líders en determinades èpoques o estacions de l'any.

2.3. ANÀLISI DE LES FONTS DE DADES

El nostre client, GLTF, ens va proporcionar una sèrie de fitxers de text que contenen les dades de les que disposa sobre l'estudi a realitzar. Aquests fitxers són els següents:

Fitxer	Format	Codificació
Equipaments.csv	CSV	UTF-8
Equipaments.xls	Excel	ANSI

Infraestructura turística.xls	Excel	ANSI
Pernotaciones.csv	CSV	UTF-8
Policies locals.xls	Excel	ANSI

A continuació procedirem a l'anàlisi general del contingut de cada fitxer.

2.3.1. Fitxer Equipaments.csv

Com és habitual en un fitxer CSV els valors estan separats per comes [,]. La primera fila correspon als noms dels camps.

Equipaments.csv		
Camp	**Descripció**	**Característiques destacables**
nom	Denominació de l'equipament en qüestió.	
adreca	Nom de la via i el número.	Tot el valor contingut entre cometes dobles ["]. No inclou la denominació de la via.
municipi	Nom del municipi al que pertany l'equipament.	
cp	Codi postal.	Hi ha valors amb menys de 5 dígits, ja que no inclouen el 0 davant.
comarca	Denominació de la comarca a la que pertany el municipi.	
telefon	Número de telèfon d'equipament.	Existeixen espais entre els números. Alguns registres no tenen telèfon.
longitud	Posició geogràfica – longitud.	
latitud	Posició geogràfica – latitud.	
categories	Diferents categories a les que pertany l'equipament.	Organització de forma jeràrquica, sent el primer terme *Equipaments,* seguit de les categories de superior a inferior. El separador entre valors del camp és – espai, barra, espai [|].
Location	Latitud i longitud d'equipament en format XML.	Els valors de latitud i longitud estan entre cometes dobles ["] i text <Point><coordinates> i ,0.0</coordinates></Point>

El fitxer *Equipaments.xls* presenta exactament les mateixes dades, però variant el format del CSV a l'Excel. Una altra diferència és que els camps *adreca* i *Location* no estan entre cometes dobles (").

2.3.2. Fitxer Infraestructura turística.xls

Es tracta d'un fitxer que engloba 18 fulls de càlcul i un d'Índex que adjunto a continuació.

Full	**Taula**	**Àmbit**	**Any**
1	Establiments hotelers	Catalunya	2013
2	Establiments hotelers	Catalunya	2012
3	Establiments hotelers	Catalunya	2011
4	Places hoteleres	Catalunya	2013

5	Places hoteleres	Catalunya	2012
6	Places hoteleres	Catalunya	2011
7	Establiments de càmpings	Catalunya	2013
8	Establiments de càmpings	Catalunya	2012
9	Establiments de càmpings	Catalunya	2011
10	Places de càmpings	Catalunya	2013
11	Places de càmpings	Catalunya	2012
12	Places de càmpings	Catalunya	2011
13	Establiments de turisme rural	Àger	2013
14	Establiments de turisme rural	Àger	2012
15	Establiments de turisme rural	Àger	2011
16	Places de turisme rural	Àger	2013
17	Places de turisme rural	Àger	2012
18	Places de turisme rural	Àger	2011

Es pot observar que comprèn les dades dels anys 2011, 2012 i 2013 en l'àmbit català del número d'establiments i el número de places d'aquests establiments.

Els tres tipus d'establiments (hotel, càmping i turisme rural) tenen plantilles diferents, però aquestes són les mateixes entre el número d'establiments i el número de places.

A continuació s'exposa l'estructura de cada tipus.

Hotels, fulls de l'1 al 6		
Cel·la	**Descripció**	**Característiques destacables**
A1	Defineix si es tracta d'establiments o places.	Discriminació es fa en la primera paraula.
A2	Subtítol.	L'última paraula abans del punt [.] indica l'any.
Columna A8-954	Nom del municipi	En format [*Nom*, *Article*]. Alguns noms no tenen article.
Columna B8-954	1 estrella	Números enters.
Columna C8-954	2 estrelles	Números enters.
Columna D8-954	3 estrelles	Números enters.
Columna E8-954	4 estrelles	Números enters.
Columna F8-954	5 estrelles	Números enters.
Columna G8-954	Total hotels amb estrelles per municipi.	Números enters.
Columna H8-954	Hostals i pensions.	Números enters.
Columna I8-954	Total hotels per municipi.	Números enters.
B956	Total català 1 estrella.	Número enter.
C956	Total català 2 estrelles.	Número enter.
D956	Total català 3 estrelles.	Número enter.

E956	Total català 4 estrelles.	Número enter.
F956	Total català 5 estrelles.	Número enter.
G956	Total hotels.	Número enter.
H956	Total hostals i pensions.	Número enter.
I956	Total global.	Número enter.

Càmpings, fulls del 7 al 12		
Cel·la	**Descripció**	**Característiques destacables**
A1	Defineix si es tracta d'establiments o places.	Discriminació es fa en la primera paraula.
A2	Subtítol.	L'última paraula abans del punt [.] indica l'any.
Columna A6-952	Nom del municipi.	En format [Nom, Article]. Alguns noms no tenen article.
Columna B6-952	Luxe.	Números enters.
Columna C6-952	Primera.	Números enters.
Columna D6-952	Segona.	Números enters.
Columna E6-952	Tercera.	Números enters.
Columna F6-952	Total càmpings per municipi.	Números enters.
B954	Total luxe.	Número enter.
C954	Total primera.	Número enter.
D954	Total segona.	Número enter.
E954	Total tercera.	Número enter.
F954	Total global	Número enter.

Turisme rural, fulls del 12 al 18		
Cel·la	**Descripció**	**Característiques destacables**
A1	Defineix si es tracta d'establiments o places.	Discriminació es fa en la primera paraula.
A2	Subtítol.	L'última paraula abans del punt [.] indica l'any.
Columna A7-953	Nom del municipi.	En format [Nom, Article]. Alguns noms no tenen article.
Columna B7-953	Casa de poble compartida.	Números enters.
Columna C7-953	Casa de poble independent.	Números enters.
Columna D7-953	Masia.	Números enters.
Columna E7-953	Masoveria.	Números enters.
Columna F7-953	Total turisme rural per municipi.	Números enters.
B955	Total cases de poble compartides.	Número enter.
C955	Total cases de poble independents.	Número enter.
D955	Total masies.	Número enter.

E955	Total masoveries.	Número enter.
F955	Total global.	Número enter.

2.3.3. Fitxer Pernoctaciones.csv

El fitxer recull les dades dels (anys 2012-2014) del número de pernoctacions i de viatgers en diferents marques turístiques de Catalunya, classificats segons si són residents a Espanya o a l'estranger. Les dades estan separades per un tabulador.

Degut a la complicada estructura del fitxer i a la manca de dades que presenta, s'ha optat per tornar a extreure les dades de l'INE[5] en format Excel. Per facilitar la comprensió i garantir l'exactitud de les dades, aquestes s'han extret, per any, dels apartats següents:

- Encuesta de ocupación hotelera [*Any*] -> Datos por zonas turísticas. Demanda -> Viajeros entrados por zonas turísticas y meses.
- Encuesta de ocupación hotelera [*Any*] -> Datos por zonas turísticas. Demanda -> Pernoctaciones de los viajeros por zonas turísticas y meses.

D'aquesta manera s'han extret 6 fitxers diferents:

Fitxer	**Format**	**Codificació**
Pernotaciones2011.xls	Excel	ANSI
Pernotaciones2012.xls	Excel	ANSI
Pernotaciones2013.xls	Excel	ANSI
Viajeros2011.xls	Excel	ANSI
Viajeros2012.xls	Excel	ANSI
Viajeros2013.xls	Excel	ANSI

Tots els fitxers estan en castellà i segueixen el mateix patró.

Pernoctaciones[*Any*].xls i Viajeros[*Any*].xls		
Cel·la	**Descripció**	**Característiques destacables**
Columna A9-18	Nom de la marca turística.	En format [*ZONA: Marca*]. Alguns al final i entre parèntesis tenen la província a la que pertanyen.
Fila B8-M	Noms dels mesos.	En format text.
Quadrant B9-M18	Dades numèriques, per marca turística i mesos.	Números enters.

Les dades en aquests fitxers presenten un canvi molt important – a partir de l'any 2012 les marques Costa Barcelona-Maresme i Costa del Garraf es fusionen en una sola marca anomenada Costa Barcelona.

2.3.4. Fitxer Policies locals.xls

El fitxer es composa de 3 fulls de càlcul i un d'Índex que adjunto a continuació.

[5] Instituto Nacional de Estadística. Hostelería y turismo. Encuesta de ocupación hotelera.

Full	Taula	Àmbit	Any
1	Efectius de les policies locals. Per graduació	Catalunya	2013
2	Efectius de les policies locals. Per graduació	Catalunya	2012
3	Efectius de les policies locals. Per graduació	Catalunya	2011

Es pot observar que comprèn les dades dels anys 2011, 2012 i 2013 en l'àmbit català del número d'efectius de policies locals.

Cada full correspon a un any diferent, presentant la següent estructura:

Policies locals.xls		
Cel·la	**Descripció**	**Característiques destacables**
A2	Subtítol.	L'última paraula abans del punt [.] indica l'any.
Columna A7-953	Nom del municipi.	En format [Nom, Article]. Alguns noms no tenen article.
Columna B7-953	Super-intendent.	Números enters.
Columna C7-953	Intendent major.	Números enters.
Columna D7-953	Intendent.	Números enters.
Columna E7-953	Inspector.	Números enters.
Columna F7-953	Sort-inspector.	Números enters.
Columna G7-953	Sergent.	Números enters.
Columna H7-953	Caporal.	Números enters.
Columna I7-953	Agent.	Números enters.
Columna J7-953	Total policies per municipi.	Números enters.
Columna K7-953	Rati de policies per mil habitants.	Números decimals.
B955	Total super-intendents.	Número enter.
C955	Total intendents major.	Número enter.
D955	Total intendents.	Número enter.
E955	Total inspectors.	Número enter.
F955	Total sort-inspectors.	Número enter.
G955	Total sergents.	Número enter.
H955	Total caporals.	Número enter.
I955	Total agents.	Número enter.
J955	Total policies.	Número enter.
K955	Total rati.	Número decimal.

Amb totes les dades mencionades podem poblar pràcticament la totalitat del nostre DW. Només ens falten dades de població, dades que relacionin les marques turístiques amb les comarques i dades sobre la dimensió temps.

3. DISSENY

3.1. DISSENY MULTIDIMENSIONAL DE LA BASE DE DADES

3.1.1. DISSENY CONCEPTUAL

Seguidament es desenvolupa l'esquema de base de dades que permet fer una descripció a alt nivell, independent del BDMS que s'utilitzi.

3.1.1.1. Els fets

Els fets (F) constitueixen el focus central d'anàlisi, allò que es vol analitzar. Totes les dades o mesuraments disponibles en el sistema han de ser organitzats en funció de cada fet. Per tant de la seva elecció depèn tot el disseny posterior.

En el nostre cas tenim diverses dades disperses. L'enunciat ens dóna la primera pista, parla del problema del desconeixement de la relació dels elements que caracteritzen un municipi i les pernoctacions que s'hi esdevenen; en definitiva, s'exposa la voluntat de conèixer els criteris decisoris dels turistes per quedar-se en un lloc o un altre.

Seguint els anàlisis posteriors queda clar que totes les dades giren al voltant de dos fets globals:

- Indicadors relacionats amb la infraestructura d'un lloc determinat.
- Indicadors relacionats amb les pernoctacions.

Així el treball està enfocat sobre dues estrelles:

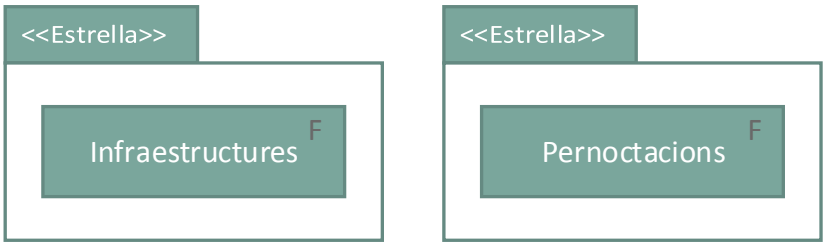


Figura 3. Estrelles principals

3.1.1.2. El gràNul escaient

El gràNul escaient és la cel·la més petita que es desitja tenir disponible. És important perquè determina la dimensionalitat de la base de dades. Un gràNul massa petit desbordaria la grandària de la base de dades, per contra un gràNul massa gran reduiria la precisió de les dades i podria implicar la renúncia d'alguna dimensió o possibilitat de calcular certes mesures.

Hi ha un gràNul per cada fet. La seva definició pels establiments és:

Característiques anuals d'una infraestructura.

I per pernoctacions:

Pernoctacions mensuals en un municipi.

3.1.1.3. Les dimensions, els seus atributs i la jerarquia interna

Un dimensió (D) representa un punt de vista específic des del qual volem analitzar les dades. Totes les dimensions van lligades al fet central de les nostres estrelles.

Dins de cada dimensió es troben diferents grups d'instàncies. Les instàncies de la mateixa granularitat s'anomenen Nivells (N). Cada nivell pot agrupar-se en un altre de més gran, formant una jerarquia d'agregació.

El primer conjunt de dimensions s'extreu de manera immediata de la definició del gràNul escaient. Podem notar que els dos fets estan relacionats amb dues dimensions – el Temps i la Zona.

- El Temps representa una dimensió bàsica en pràcticament qualsevol model dimensional. Ajuda a temporitzar totes les dades en unitats de temps predefinides. En el nostre cas la temporització es fa a nivell de mes, estació, trimestre i any.

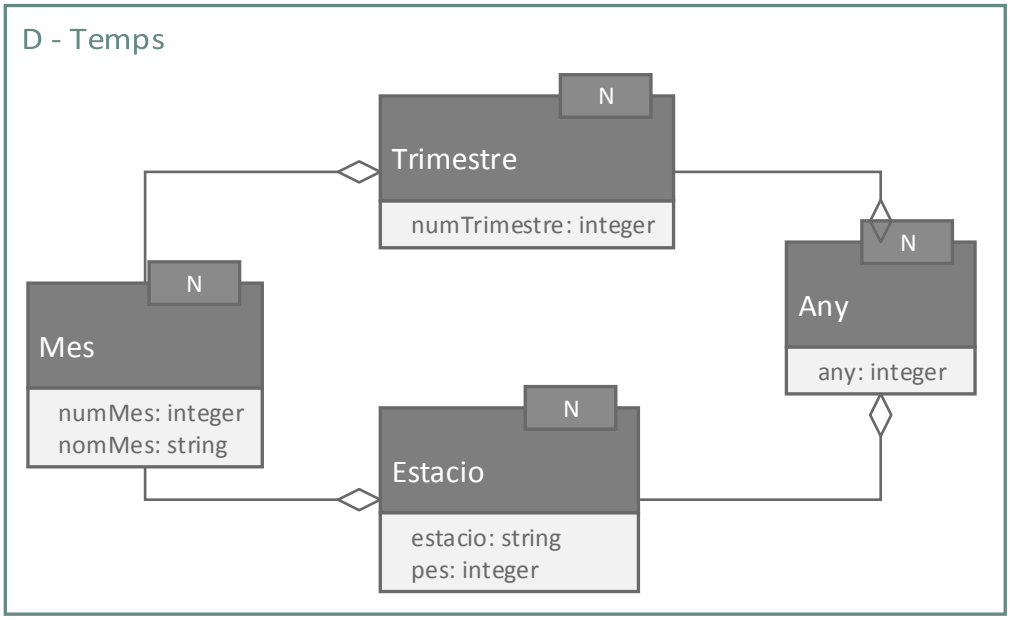


Figura 4. Dimensió Temps

- La Zona representa el lloc geogràfic de l'estudi. S'organitza de manera jeràrquica a partir dels municipis, cap a comarca, marca turística i comunitat autònoma.

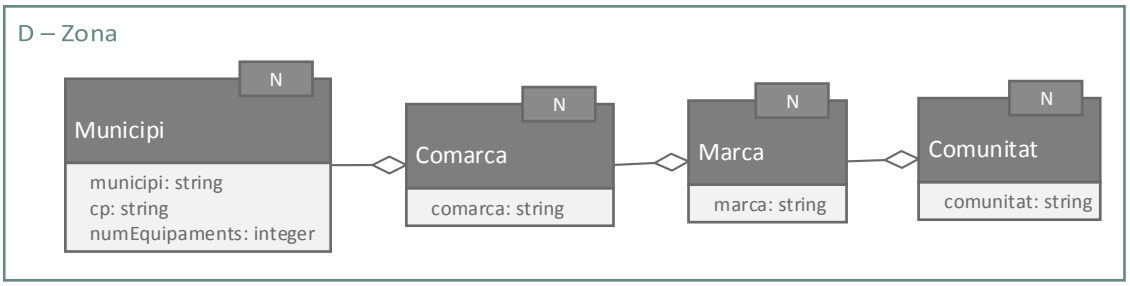


Figura 5. Dimensió Zona

Però també hi ha altres dimensions que es deriven directament de la Zona. Aquest és el cas d'Equipament i Establiment. Tornant als requeriments exposats en l'anàlisi detallat, podem observar que el concepte de la dimensió Establiment sempre es demana en una franja temporal, en altres paraules, sempre es demana en el fet Infraestructures. Per aquesta raó s'ha fet un trasllat d'aquesta dimensió cap al fet Infraestructures.

- Els Equipaments es classifiquen segons la funcionalitat que ofereixen per tipus i subtipus, estant els subtipus inclosos dins del tipus de manera jeràrquica.

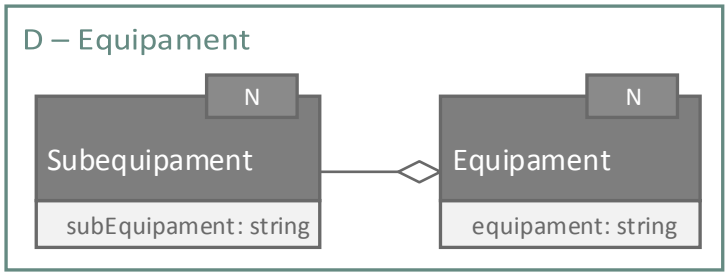


Figura 6. Dimensió Equipament

- Els Establiments hotelers els agrupem en 3 tipus diferents – hotels, càmpings i allotjaments rurals.

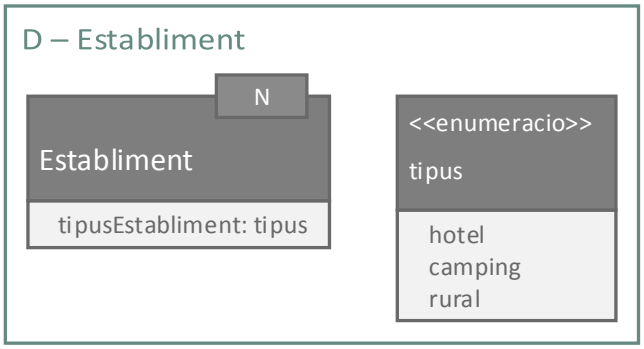


Figura 7. Dimensió Establiment

D'aquesta manera tenim definit el primer disseny general de les nostres estrelles.

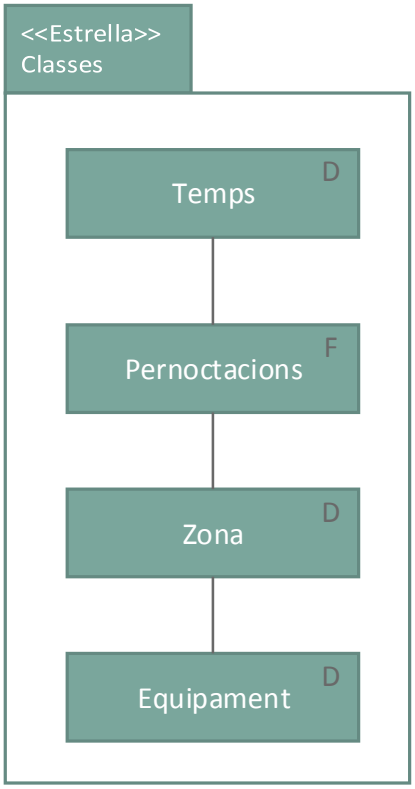


Figura 8. Estrella Pernoctacions

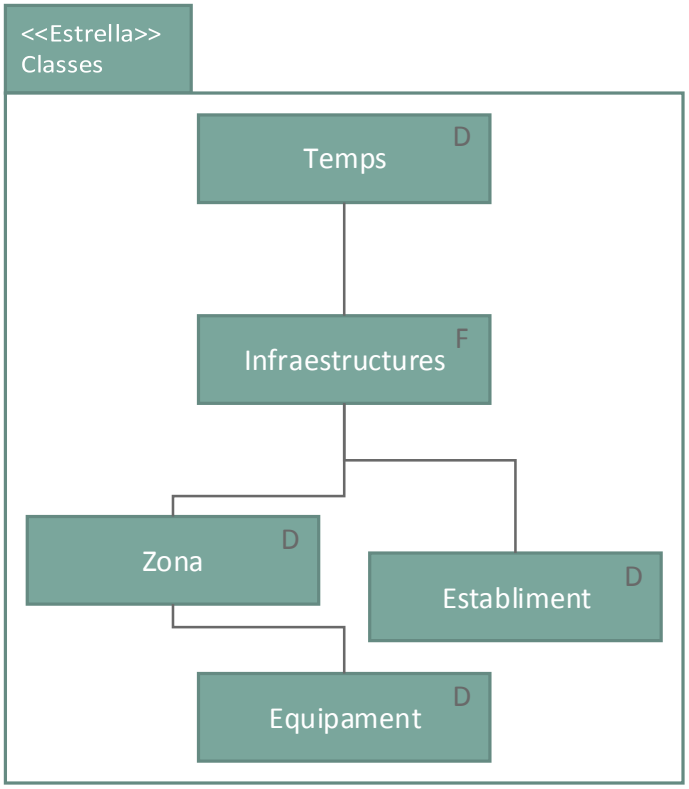


Figura 9. Estrella Infraestructures

3.1.1.4. Mesures dels fets i Cel·les

Les mesures són els atributs numèrics (normalment additius) que resulten de la connexió de totes les dimensions dins del fet.

En el nostre cas tenim diferents tipus de mesures: referents a la zona, als equipaments, establiments i a les pernoctacions.

- Referents a la zona: cens municipal, número d'efectius policials.
- Referents als establiments: número d'establiments, places de cada tipus d'establiment.
- Referents a les pernoctacions: número de pernoctacions, número de viatgers.

Totes les mesures esmentades no tenen registres unitaris, sinó que es sumen mensualment segons el municipi i el tipus contret, o sigui, segons la nostra granularitat mínima. Aquestes dades s'entraran posteriorment ja calculades dins del magatzem de dades.

Pel que es refereix a les Cel·les, aquestes es creen segons les diferents granularitats presents en cada Fet. Com que les granularitats són úniques per cada fet, només disposem d'una Cel·la per cada Fet.

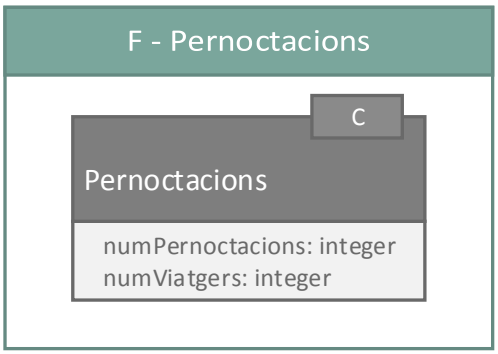


Figura 10. Fet Pernoctacions

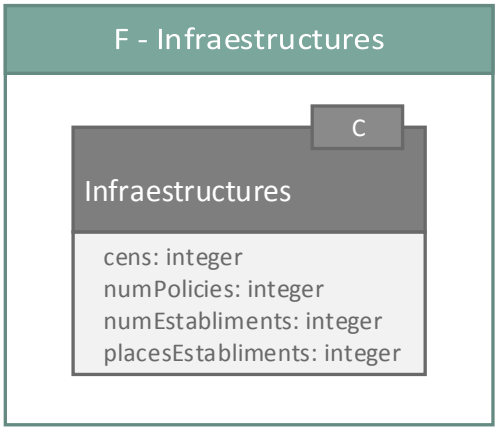


Figura 11. Fet Infraestructures

3.1.1.5. Restriccions d'integritat

Hi ha tres tipus de restriccions d'integritat: bases, restriccions d'agregació i transitivitat. Les Bases són els diferents conjunts de Nivells que poden definir una Cel·la inequívocament. La Cel·la Pernoctacions es pot definir a partir del Mes i la Marca. Per l'altra banda, la Cel·la Infraestructures es defineix a partir de l'Any, Municipi, Equipament i Establiment.

Pel que fa a les restriccions d'agregació cal destacar vàries coses. En primer lloc, totes les operacions de l'estrella de les Pernoctacions són compatibles, disjuntes i completes. Per tant no hi ha cap problema per efectuar operacions d'agregació o transitivitat. En segon lloc, les operacions d'agregació en el temps de les mesures de les Infraestructures s'han de fer operant amb la mitjana, ja que es registra una quantitat total cada cert temps. La resta de restriccions es compleixen igual que en l'apartat anterior.

El model conceptual final queda de la següent manera:

(pg. següent)

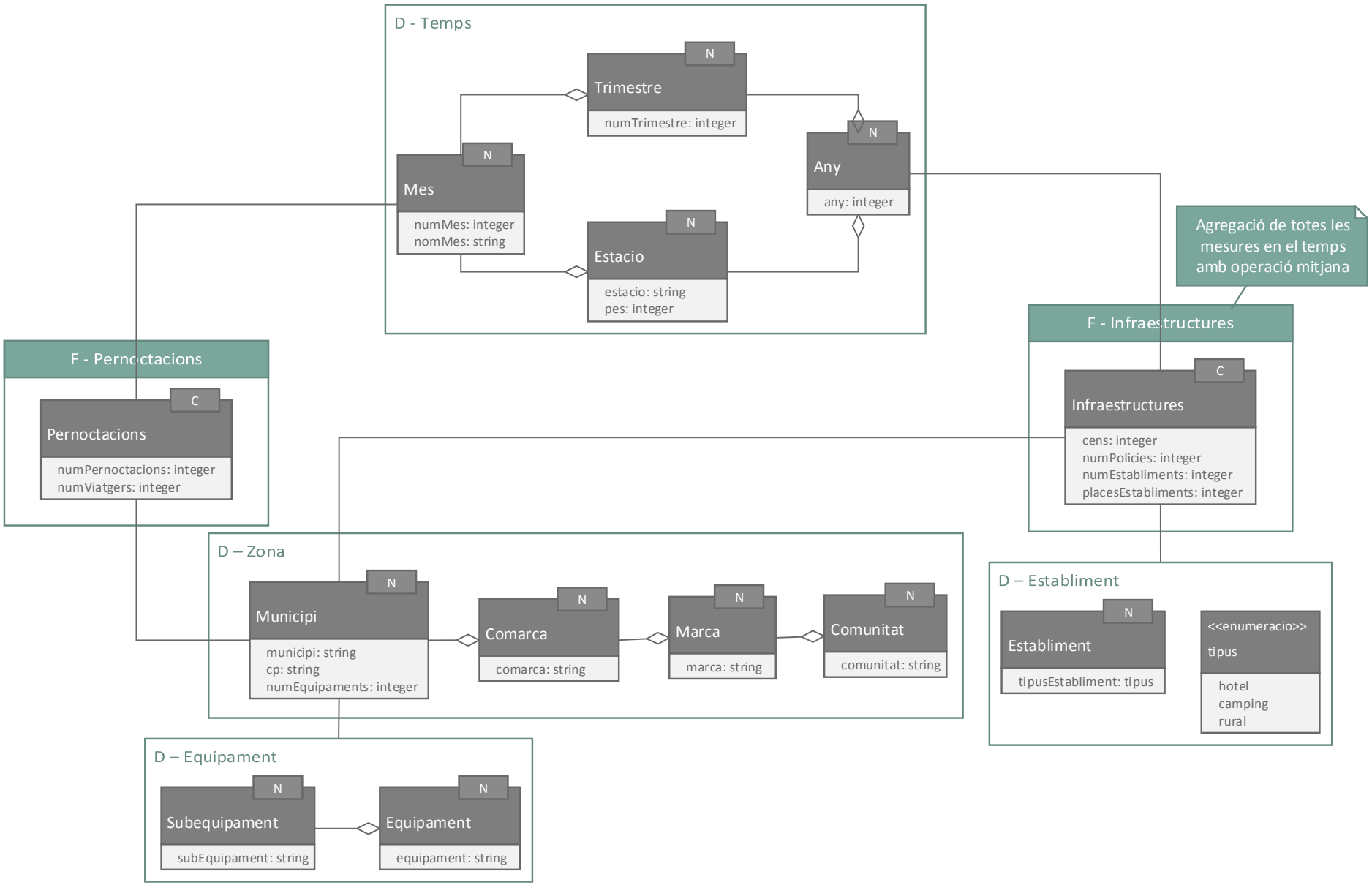


Figura 12. Model conceptual de la base de dades

3.1.1.6. Estudi de la viabilitat

Per estimar la grandària de la nostra base de dades, ens centrem en els Fets, ja que el volum d'aquestes taules representa pràcticament la totalitat del volum de la base de dades. Les Dimensions, en comparació, ocupen un espai insignificant.

Com que en aquest punt encara no sabem l'organització exacta de les claus que representen els Fets, s'agafarà una estimació màxima, a partir de les màximes bases possibles. Tindrem present en el càlcul que el temps d'estudi és de 3 anys i que la zona geogràfica està acotada a Catalunya.

Primer calculem el Fet Pernoctacions a partir de la Base {Mes, Municipi}. Sabem que hi ha 12 mesos a l'any i que segons Idescat[6] a Catalunya hi ha 943 municipis, que arrodonirem a 1.000. Per tant tenim un espai de 12 x 1.000 x 3 = 36.000 possibles cel·les.

L'espai d'una cel·la de Pernoctacions es basa en claus primàries referents a enters de marca (4 bytes) i mes (4 bytes) més dos atributs enters de 4 bytes. D'aquesta manera una cel·la ocupa 4 x 4 = 16 bytes, sent l'espai estimat per Pernoctacions de 16 x 36.000 ≈ 0,55 MB.

Ara calculem el Fet Infraestructura a partir de la Base {Any, Municipi, Establiment}. Sabem que treballem en un espai temporal de 3 anys, i considerem 1.000 municipis diferents. Com a molt en un municipi hi haurà una cinquantena d'establiments diferents. Per tant tenim un espai de 3 x 1.000 x 50 = 150.000 possibles cel·les.

L'espai d'una cel·la d'Infraestructura es basa en claus primàries referents a enters d'any (4 bytes), municipi (4 bytes), establiment (4 bytes), més 4 atributs enters de 4 bytes. D'aquesta manera una cel·la ocupa 4 x 7 = 28 bytes i l'espai estimat per Infraestructura és de 28 x 150.000 ≈ 4 MB.

Com podem veure els dos fets sumen 0,55 + 4 ≈ 4,55 MB, un espai prou petit. Hem de considerar que les dimensions la base de dades no haurien de créixer més de 10 MB.

3.1.2. DISSENY LÒGIC

El següent pas en el disseny multidimensional consisteix en la traducció del model conceptual al model lògic. Les dades es guarden en un SGBD relacional, pel que utilitzarem una implementació ROLAP.

[6] Institut d'Estadística de Catalunya, consulta de municipis de l'any 2013.

Tot el disseny conceptual s'ha fet seguint el model estrella. Els UMLs expressen tots els nivells que poden haver-hi, de manera que es pot caure fàcilment en l'error de realitzar la traducció al disseny lògic totalment normalitzat, o sigui, realitzar un floc de neu. Si bé aquest tipus de dissenys són molt vàlids per bases de dades operacionals, en el nostre cas suposaria una considerable pèrdua de rendiment, amb un guany d'espai normalment insignificant. Per aquest motiu s'ha realitzat un anàlisi dimensió per dimensió i fet per fet buscant un equilibri entre l'espai i el rendiment.

Abans de descriure l'anàlisi, vull apuntar que en totes les taules creades s'utilitza per defecte com a clau primària, una clau substituta entera i auto-incremental. Aquest fet dóna dos avantatges essencials:

- Permet reduir la grandària de les claus primàries de les dimensions i pel consegüent la grandària de les taules Fet que referencien aquestes dimensions.
- Serveix com a mesura de contenció en els possibles canvis posteriors dels atributs identificadors de cada fila.

3.1.2.1. Dimensió Temps

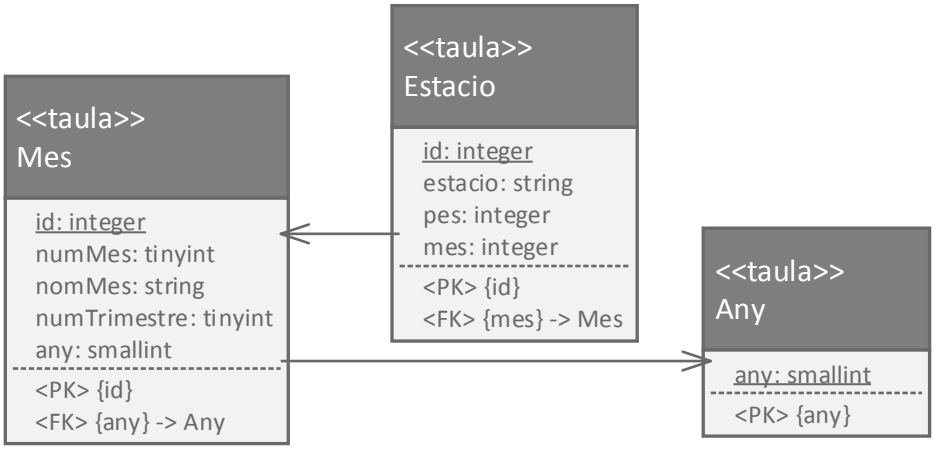


Figura 13. Dimensió Temps

Aquesta dimensió és la més petita i inclou quatre nivells: Mes, Trimestre, Estació i Any. Tot són instàncies heterogènies representades via especificació de la dimensió.

En els dos extrems hi ha el Mes i Any que estan relacionades amb fets diferents, o sigui, fan que la granularitat no sigui uniforme. Això implica que en les consultes, l'operació JOIN de cada fet sigui més o menys complexa (multiplica una quantitat de dades diferents). Comptant que les consultes més freqüents de cada fet es faran segons la seva granularitat assignada es

considera que és millor opció separar els nivells Mes i Any en dues taules diferents.

A més a més, la taula Any a priori només guardarà 3 dades – 2011, 2012 i 2013. Aquí no hi pot haver confusió o canvi d'atribut identificatiu. La creació d'una clau primària substituta es converteix en un fet innecessari, que només complica les consultes, ja que obliga a fer el JOIN entre Mes i Any quan es vulgui saber a quin any pertany un determinat mes. Per aquest motiu en la taula Any la clau primària serà el propi número de l'any.

Pel que fa als trimestres, aquests es poden integrar a la taula de Mes. No obstant, no es pot fer el mateix amb les estacions, ja que aquests no són exclusives entre si per mesos. Si optéssim per la mateixa estratègia, duplicaríem innecessàriament els mesos, complicant la gestió i augmentant molt la grandària de les Pernoctacions. D'aquí es deriva una nova taula Estació.

3.1.2.2. Dimensió Zona

Figura 14. Dimensió Zona

En aquesta dimensió, igual que en l'anterior, ens trobem amb un ordre jeràrquic de nivells. Sabem que el nivell Municipi es relaciona directament amb dos fets, per tant aquest ha de formar una taula.

Els nivells jeràrquics superiors, incloent Comarca, Marca i Comunitat tenen realment poques dades. Per aquesta raó i prioritzant el rendiment s'integra tots aquests nivells dins de la taula Municipi.

3.1.2.3. Dimensió Establiment

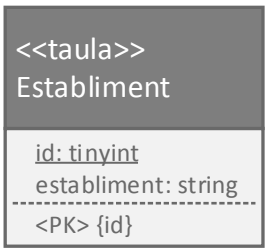


Figura 15. Dimensió Establiment

Aquesta dimensió és molt petita, només conté 3 dades. Per facilitar-ne la gestió s'implementa com una taula amb diferents valors.

Es podria seguir la mateixa estratègia que s'ha utilitzat en la taula Any, no utilitzar una clau primària substituta, ja que hi ha molt pocs valors. Ara bé, en aquest cas l'atribut identificatiu és una cadena de text, que pot ocupar molt més espai que en el cas d'any i per consegüent farà créixer encara més el fet Infraestructures, ja prou gran per si mateix. També cal ressaltar que l'atribut establiment és susceptible a modificacions posteriors, pel que és convenient poder-lo editar sense afectar la totalitat del sistema.

3.1.2.4. Dimensió Equipament

Figura 16. Dimensió Equipament

En la dimensió Equipament trobem dos nivells organitzats jeràrquicament. Per optimitzar el rendiment s'engloben en una sola taula, cada fila de la qual representa una suma dels equipaments i sub-equipaments en un municipi, per tipus.

3.1.2.5. Fet Pernoctacions

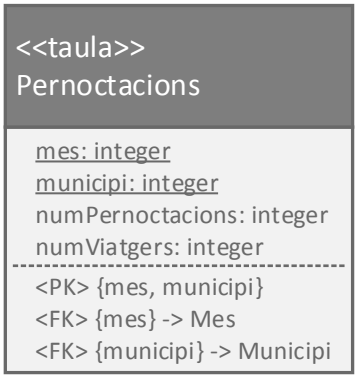


Figura 17. Fet Pernoctacions

El fet Pernoctacions conté només una Cel·la que s'identifica pel mes i el municipi. Com que aquesta taula no estarà referenciada per cap altra, prescindirem de la creació d'una clau primària substituta.

3.1.2.6. Fet Infraestructures

Figura 18. Fet Infraestructures

Igual que en el cas anterior, aquest fet conté només una Cel·la que s'identifica per any, municipi, establiment i equipament. En aquest cas tampoc es crea una clau primària substituta.

A continuació es presenta la totalitat del model lògic.

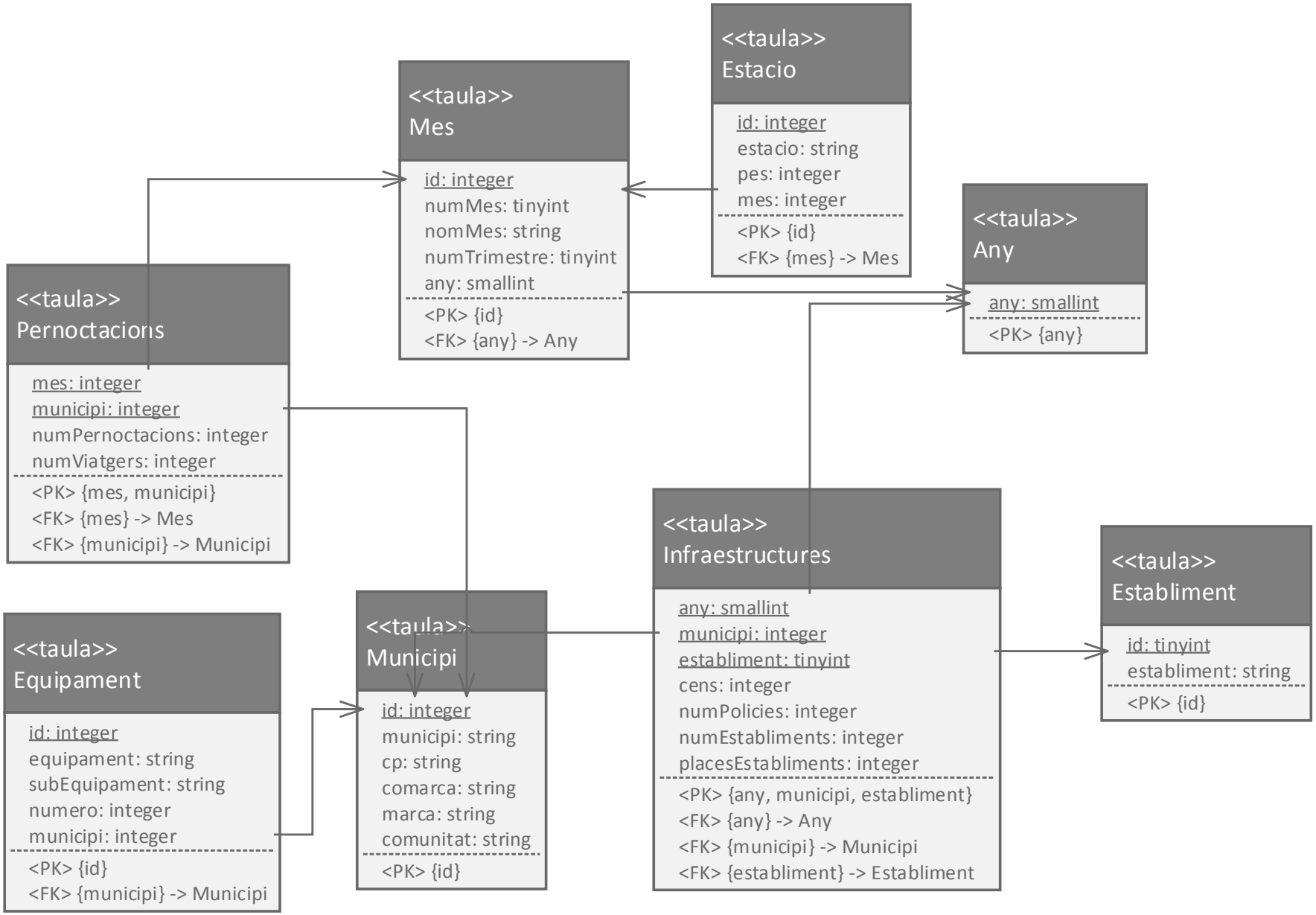


Figura 19. Model lògic de la base de dades

3.2. DISSENY DELS PROCESSOS ETL

Com és habitual se'ns han lliurat una sèrie d'arxius de text, en diferents formats que contenen les dades de les que disposa el client. Està clar que aquestes dades no es poden carregar directament al nostre DW, primer han de passar per un procés ETL. Aquest procés rau entre les fonts de dades i el DW, i com indica el seu nom segueix els passos d'extracció, transformació i càrrega de dades. Tot el procés està pensat per ser cíclic, en altres paraules, pot executar-se cada cert temps.

El primer pas, extracció, s'encarrega d'extreure les dades de les seves fonts en un dels formats llegibles i estàndards. També assegura que compleixin els requeriments demanats. Una vegada extretes es prepara el seu tractament posterior.

El segon pas, transformació, recull les dades extretes i aplica una sèrie de regles i funcions sobre aquestes per poder-les carregar en el DW. Entre aquestes funcions destaquen: neteja, filtratge, re-ordenament, càlculs, traduccions, agregacions, etc. Totes aquestes funcions han d'adaptar i validar les dades per a que aquestes puguin ser enviades a l'últim pas.

Per últim, tenint les dades preparades es fa una càrrega d'aquestes al destí – la nostra base de dades, assegurant la correctesa del procés i resolent qualsevol conflicte.

3.2.1. EXTRACCIÓ

El client ens ha proporcionat la majoria de dades, hem analitzat en l'apartat d'anàlisi. No obstant hi ha altres dades que s'han hagut de buscar per mitjans propis. Aquest és el cas de les dades de població, dades que relacionin les marques turístiques amb les comarques i sobre la dimensió temps.

3.2.1.1. Població

Les dades sobre la població i en especial en cens municipal s'han extret d'INE[7]. Concretament des del següent apartat:

- Demografía y población -> Población de municipios y unidades poblacionales -> Cifras oficiales de población resultantes de la revisión del Padrón municipal a 1 de enero -> Detalle municipal

Una vegada descomprimit, ens interessen els fitxers corresponents als anys 2011, 2012 i 2013:

[7] Instituto Nacional de Estadística. Archivo comprimido con los ficheros Excel municipales de cada año.

Fitxer	Format	Codificació
pobmun11.xls	Excel	ANSI
pobmun12.xls	Excel	ANSI
pobmun13.xls	Excel	ANSI

El format dels fitxers és el següent:

pobmun[*Any-2000*].xls		
Cel·la	**Descripció**	**Característiques destacables**
A1	Títol.	L'última paraula indica l'any.
Columna A3-8118	Codi província.	En format text.
Columna B3-8118	Província.	En format text.
Columna C3-8118	Codi del municipi.	En format text.
Columna D3-8118	Nom oficial del municipi.	En format text.
Columna E3-8118	Total població per municipi.	Número enter.
Columna F3-8118	Total homes per municipi.	Número enter.
Columna G3-8118	Total dones per municipi.	Número enter.

3.2.1.2. Relació comarques – marques turístiques

Com que no s'ha pogut trobar un fitxer fiable que contingui aquesta relació s'ha decidit elaborar una taula pròpia a partir de la informació extreta d'Agència Catalana de Turisme[8].

Figura 20. Mapa de marques turístiques catalanes

[8] Agència Catalana de Turisme. Dossier de premsa 2014.

Marca turística	Comarques que conté
Barcelona	Barcelonès
Catalunya Central	Anoia Bages Osona
Costa Barcelona	Alt Penedès Baix Llobregat Garraf Maresme Vallès Occidental Vallès Oriental
Costa Brava	Alt Empordà Baix Empordà Gironès Pla de l'Estany Selva
Costa Daurada	Alt Camp Baix Camp Baix Penedès Conca de Barberà Priorat Tarragonès
Pirineus	Alt Urgell Alta Ribagorça Berguedà Cerdanya Garrotxa Pallars Jussà Pallars Sobirà Ripollès Solsonès
Terres de Lleida	Garrigues Noguera Pla d'Urgell Segarra Segrià Urgell
Terres de l'Ebre	Baix Ebre Montsià Ribera d'Ebre Terra Alta
Vall d'Aran	Val d'Aran

3.2.1.3. Temps

La dimensió Temps és relativament petita i molt matemàtica. La taula anys només conté 3 dades i la taula mesos conté els mesos i trimestres corresponents a aquests anys. Per aquesta raó s'ha pres la decisió d'omplir aquestes taules manualment amb els següents valors:

Any
2011
2012
2013

Mes			
Número del Mes	**Nom del Mes**	**Número del trimestre**	**Any**
1	gener	1	2011
2	febrer	1	2011
3	març	1	2011
4	abril	1	2011
5	maig	2	2011
6	juny	2	2011
7	juliol	2	2011
8	agost	2	2011
8	setembre	3	2011
10	octubre	3	2011
11	novembre	3	2011
12	desembre	3	2011
1	gener	1	2012
2	febrer	1	2012
3	març	1	2012
4	abril	1	2012
5	maig	2	2012
6	juny	2	2012
7	juliol	2	2012
8	agost	2	2012
8	setembre	3	2012
10	octubre	3	2012
11	novembre	3	2012
12	desembre	3	2012
1	gener	1	2013
2	febrer	1	2013
3	març	1	2013
4	abril	1	2013
5	maig	2	2013
6	juny	2	2013
7	juliol	2	2013
8	agost	2	2013
8	setembre	3	2013
10	octubre	3	2013
11	novembre	3	2013
12	desembre	3	2013

Pel que es refereix a les diferents estacionalitats, no s'ha trobat cap informació en format estàndard que relacioni estacions –turístiques- de l'any amb els seus mesos. Per aquesta raó s'ha decidit, a partir del calendari, omplir la taula d'estacions amb informació pròpia. A més a més, la base de dades està preparada per ampliar les estacions de manera fàcil, per quan es necessiti.

Estació	**Freqüència**	**Pes**
hivern	Tots els anys en els mesos 12, 1 i 2 consecutius.	1
primavera	Tots els anys en els mesos 3, 4 i 5 consecutius.	1
estiu	Tots els anys en els mesos 6, 7 i 8 consecutius.	1

tardor	Tots els anys en els mesos 9, 10 i 11 consecutius.	1
festes nadalenques	Tots els anys en els mesos 12 i 1 consecutius.	1
pasqua	Els anys 2011 i 2012 ens el més 4 i el 2013 en el 5.	1
vacances estiu	Tots els anys en els mesos 7, 8 i 9.	1

3.2.2. TRANSFORMACIÓ

Una vegada recopilades i extretes totes les dades de les seves fonts originals s'inicia el procés de tractament de les mateixes que permet netejar-les i adaptar-les a la nostra base de dades.

Anem a destacar els procediments a alt nivell per preparar les dades taula per taula.

3.2.2.1. Any, Mes i Estació

Com ja s'ha raonat, no es fa cap procés de transformació, sinó que directament es procedeix a carregar la informació detallada en l'apartat anterior directament a la base de dades.

3.2.2.2. Establiment

Es tracta d'una altra taula molt petita. Fins i tot en el disseny conceptual els valors que conté s'han considerat com una enumeració. Per aquesta raó es passarà directament a entrar la informació a la taula de manera manual. La classificació dels tipus d'establiments és la següent:

Establiment
hotel
camping
rural

3.2.2.3. Municipi

Aquesta taula guarda informació de noms de municipis, codis postals, comarques, marques turístiques i comunitat.

El camp nom de municipi al principi havia de ser extret del fitxer pobmun13.xls, ja que s'ha suposat que contindria tots els municipis de Catalunya amb els seus noms oficials actualitzats. Més endavant s'ha vist que no és així i que els noms en aquest fitxer estan molt desfasats (alguns municipis van canviar el seu nom oficial ja en any 2006, però en el fitxer seguien utilitzant els noms antics). Per aquesta raó s'ha canviat l'estratègia i s'ha decidit extreure els noms del fitxer Infraestructura turística.xls, on estan correctament actualitzats.

El codi postal i la comarca es carreguen a partir del fitxer Equipaments.csv. Les dades s'han d'homogeneïtzar i creuar amb les dades del municipi.

Pel que fa a la marca turística, la informació s'extreu del fitxer propi elaborat en l'apartat anterior. També cal homogeneïtzar les dades i creuar-les contra les comarques de cada registre.

Finalment la comunitat és la mateixa en tots els registres – Catalunya.

3.2.2.4. Equipament

La taula guarda els camps d'equipament i sub-equipament que pertanyen a un municipi i el seu nombre sumatori, per si n'hi ha més d'un igual.

Les dades s'extrauen del fitxer Equipaments.csv. Es descarten totes les dades del fitxer, menys el camp Municipi i Categories. Després es parteix el camp Categories en dos camps, corresponents al segon i tercer terme del camp (recordem que els termes estan separats per un espai, una barra i un altre espai), que en el seu torn corresponen a l'equipament i sub-equipament.

El següent pas consisteix en fer una suma, en un altre camp, dels registres que tenen iguals el Municipi, l'Equipament i Sub-equipament, eliminant els duplicats.

Finalment les dades s'homogeneïtzen, quadrant els municipis amb els entrats en la taula Municipi.

3.2.2.5. Pernoctacions

Es tracta de la primera taula que representa un Fet, per tant cal calcular les mesures. La taula registra els camps referents al número de pernoctacions i viatgers en un municipi i un mes determinat.

Ja s'ha detectat que no existeixen estudis disponibles que relacionin les pernoctacions i municipis, per tant es fa una aproximació a partir de les pernoctacions per marca turística. Aquesta aproximació consisteix en distribuir les pernoctacions segons el número de places hoteleres de cada municipi.

El primer fitxer que es necessita és Infraestructura turística.xls. D'ell s'extreu el número de places de diferents tipus d'establiments per municipi. Primer cal extreure els noms del municipis, juntament amb les places hoteleres, de càmping i turisme rural per cada any. Seguidament en una columna separada cal sumar tots els tipus d'establiments per municipi anuals.

Paral·lelament necessitem treballar els fitxers Pernoctaciones[*Any*].xls i Viajeros[*Any*].xls. De tots 6 fitxers cal extreure el número de pernoctacions i el número de viatgers per marca i mes determinat.

Evidentment les dades han de ser homogeneïtzades. Sabem que en els fitxers de pernoctacions i viatgers del 2011 hi ha una marca de més i una altra amb un nom diferent. Aquestes marques són Costa del Garraf i Costa-Maresme. Per unificar el criteri de les marques amb les dades posteriors es fa una suma de les dades del Costa del Garraf i Costa-Maresme en una sola marca, Costa Barcelona.

Finalment es procedeix a distribuir les pernoctacions i viatgers per municipis. Primer s'agrupen els registres del primer pas segons les marques (tenim la relació marca – municipi en la taula Municipi), es fa una suma de places que representa el total. S'agrega una nova columna amb el percentatge de cada municipi sobre el total de cada marca turística. El càlcul de les mesures es fa multiplicant el percentatge de cada municipi pel número de pernoctacions i separadament pel número de viatgers de cada marca turística.

3.2.2.6. Infraestructures

Aquesta és la segona taula que representa el Fet i de la que cal calcular les mesures. Aquí es guarden tots els registres relatius als municipis, anys i establiments (mesures de cens, número de policies, número d'establiments i el número de places dels establiments).

El primer pas és pràcticament el mateix que en l'apartat anterior. Del fitxer Infraestructura turística.xls s'extreu el número de places i el número total de diferents tipus d'establiments per municipi i any. Primer cal extreure i homogeneïtzar els noms del municipis, juntament amb la quantitat d'hotels i places hoteleres, el mateix per càmping i turisme rural per cada any. Queda clar que cal agrupar tots els hotels en un tipus d'establiment hotel, tots els càmpings en tipus càmping i tot el turisme rural en tipus rural.

En la segona fase l'afegeix el cens municipal a partir dels fitxers pobmun[*Any-2000*].xls. Aquests fitxers primer passen una fase de pre-processament que elimina tots els municipis menys els que pertanyen a Catalunya. Després els noms es normalitzen segons els entrats a la base de dades. Finalment s'afegeix el cens de cada municipi en el registre corresponent.

La tercera i última fase és el càlcul del número de policies per municipi i any. L'extracció de dades es fa del fitxer Policies locals.xls. En aquest fitxer ja hi ha el total de policies per cada municipi, qer tant no cal fer cap càlcul addicional. Es fa una extracció de municipis amb l'homogeneïtzació habitual juntament amb la suma de policies per cada municipi. Aquesta suma és la que s'afegeix en els registres corresponents de la taula Infraestructures.

3.2.3. CÀRREGA

Una vegada disposem de les taules temporals amb les dades que volem inserir en el format desitjat, només cal fer la càrrega d'aquestes cap a la base de dades. Per evitar possibles problemes d'integritat és important realitzar la càrrega en un ordre correcte. Aquest ordre és:

1. Any
2. Mes
3. Estació
4. Establiment
5. Municipi
6. Equipament
7. Pernoctacions
8. Infraestructures

Els passos 7 i 8 poden intercanviar l'ordre sense cap problema.

4. IMPLEMENTACIÓ

4.1. BASE DE DADES

4.1.1. ENTORN

El treball es fa amb el motor de bases de dades MySQL versió 5.1.75, que ja estava carregat en la màquina virtual, amb una configuració bàsica. La connexió local es realitza via usuari **root** i contrasenya **TFC2014**.

Una vegada connectats al DBMS, s'ha vist que ja existia una base de dades amb el nom **tfc_dw**. Com que volem tenir el control sobre aquesta des de la seva creació, l'hem tornat a crear des de zero.

```
DROP DATABASE IF EXISTS tfc_dw;
CREATE DATABASE tfc dw DEFAULT CHARACTER SET utf8;
```

Degut a que només treballem amb aquesta base de dades concreta, per comoditat és millor seleccionar-la com a base de dades per defecte.

```
USE tfc dw;
```

4.1.2. TAULES

Les taules es creen traduint directament el disseny lògic de la base de dades a un de físic adaptat al nostre DBMS. Seguidament s'exposen les instruccions adaptades al MySQL 5.1:

```
CREATE TABLE Any_ (
      any_ SMALLINT,
      CONSTRAINT pk_any_ PRIMARY KEY (any_)
);
CREATE TABLE Mes (
      id INT AUTO_INCREMENT,
      numMes TINYINT NOT NULL,
      nomMes VARCHAR(10),
      numTrimestre TINYINT,
      any_ SMALLINT,
      CONSTRAINT pk_mes PRIMARY KEY (id),
      CONSTRAINT fk_mesAny FOREIGN KEY (any_) REFERENCES Any_
(any_)
            ON DELETE RESTRICT ON UPDATE RESTRICT
);
CREATE TABLE Estacio (
      id INT AUTO_INCREMENT,
```

```
      estacio VARCHAR(30) NOT NULL,
      mes INT,
      pes INT,
      CONSTRAINT pk_estacio PRIMARY KEY (id),
      CONSTRAINT fk_estacioMes FOREIGN KEY (mes) REFERENCES Mes
(id)
);
CREATE TABLE Establiment (
      id TINYINT AUTO_INCREMENT,
      establiment VARCHAR(20) NOT NULL,
      CONSTRAINT pk_establiment PRIMARY KEY (id)
);
CREATE TABLE Municipi (
      id INT AUTO_INCREMENT,
      municipi VARCHAR(50) NOT NULL,
      cp CHAR(5),
      comarca VARCHAR(40),
      marca VARCHAR(30),
      comunitat VARCHAR(10),
      CONSTRAINT pk_municipi PRIMARY KEY (id)
);
CREATE TABLE Equipament (
      id INT AUTO_INCREMENT,
      equipament VARCHAR(50) NOT NULL,
      subEquipament VARCHAR(100),
      numero INT,
      municipi INT NOT NULL,
      CONSTRAINT pk_equipament PRIMARY KEY (id),
      CONSTRAINT fk_equipamentMunicipi FOREIGN KEY (municipi)
REFERENCES Municipi (id)
);
CREATE TABLE Pernoctacions (
      mes INT,
      municipi INT,
      numPernoctacions INT,
      numViatgers INT,
      CONSTRAINT pk_pernoctacions PRIMARY KEY (mes, municipi),
      CONSTRAINT fk_pernoctacionsMes FOREIGN KEY (mes) REFERENCES
Mes (id) ON DELETE RESTRICT ON UPDATE RESTRICT,
      CONSTRAINT fk_pernoctacionsMunicipi FOREIGN KEY (municipi)
REFERENCES Municipi (id) ON DELETE RESTRICT ON UPDATE RESTRICT
);
CREATE TABLE Infraestructures (
      any_ SMALLINT,
      municipi INT,
      establiment TINYINT,
```

```
      cens INT,
      numPolicies INT,
      numEstabliments INT,
      placesEstabliments INT,
      CONSTRAINT pk_infraestructures PRIMARY KEY (any_, municipi,
establiment),
      CONSTRAINT fk_infraestructuresAny_ FOREIGN KEY (any_)
REFERENCES Any_ (any_) ON DELETE RESTRICT ON UPDATE RESTRICT,
      CONSTRAINT fk_infraestructuresMunicipi FOREIGN KEY (municipi)
REFERENCES Municipi (id) ON DELETE RESTRICT ON UPDATE RESTRICT,
      CONSTRAINT fk_infraestructuresEstabliment FOREIGN KEY
(establiment) REFERENCES Establiment (id) ON DELETE RESTRICT ON
UPDATE RESTRICT
);
```

4.1.3. USUARIS

El super-usuari ja existent **root** només ha de servir per executar modificacions en l'estructura de la base de dades o eliminacions i modificacions excepcionals de les dades. Com que es tracta d'un usuari amb tots els privilegis possibles, aquest no ha de ser utilitzat per tasques diàries i només es pot connectar des de l'ordinador local (comportament configurat per defecte).

Així, cal crear dos usuaris addicionals:

etl: usuari que s'utilitza només en el procés ETL. Té permisos de selecció, inserció i actualització de totes les taules. La seva connexió es preveu només des d'ordinador local.

normal: usuari que s'utilitza per generar informes i extreure coneixement general del DW. Només té permisos de selecció, ja que en principi no es preveu que pugui fer modificacions. La seva connexió pot ser local o remota indistintament.

Per simplificar i pensant que estem en un entorn de proves, tots els usuaris tenen la mateixa contrasenya assignada, **TFC2014**. Les instruccions de creació són les següents:

```
CREATE USER `etl`@`localhost` IDENTIFIED BY 'TFC2014';
CREATE USER `normal`@`%` IDENTIFIED BY 'TFC2014';
GRANT SELECT, INSERT, UPDATE ON `tfc_dw`.* TO
      `etl`@`localhost`;
GRANT SELECT ON `tfc_dw`.* TO `normal`@`%`;
```

Finalment tenim generada tota l'estructura necessària per guardar les dades.

4.2. ETL

4.2.1. ENTORN

El treball es fa amb la suite BI de Pentaho 5.1, concretament amb l'eina Spoon que permet automatitzar tot el cicle ETL. En la màquina virtual ja el tenim carregat en la seva configuració base. Cada vegada que creem un nou informe cal crear una nova connexió amb el DW.

Obrim l'assistent de creació d'una nova connexió (**File→New→Database Connection**) i l'omplim amb els següents paràmetres:

Paràmetre	**Valor**
Connection Name:	DW
Connection Type:	MySQL
Access:	Native (JDBC)
Host Name:	localhost
Database Name:	tfc_dw
Port Number:	3306
User Name:	etl
Password:	TFC2014

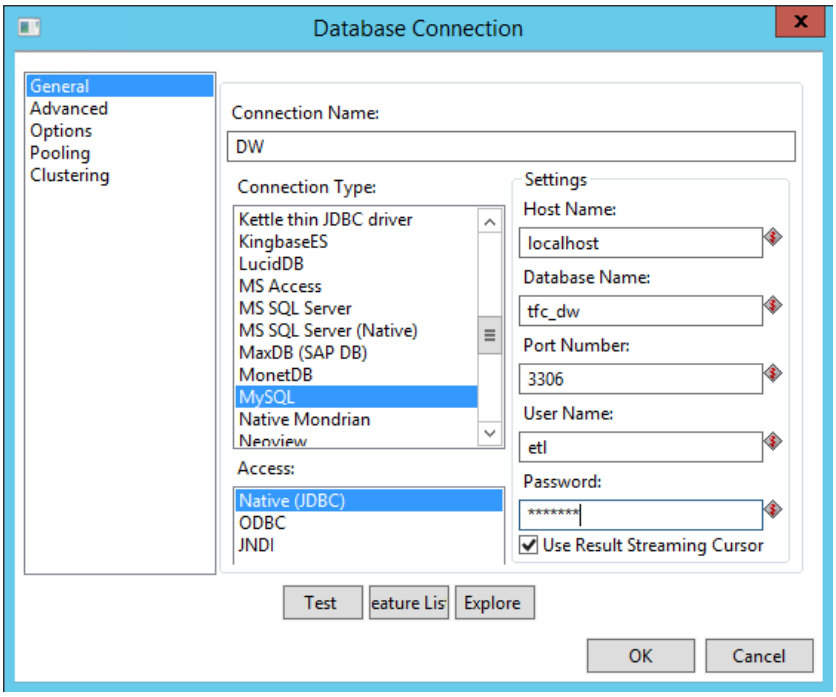


Figura 21. Connexió del Spoon amb la base de dades

La connexió amb el DW estarà identificada pel nom DW.

4.2.2. PROCESSOS AUTOMATITZATS

A continuació s'exposaran els punts més destacats de la creació de processos ETL amb Spoon.

4.2.2.1. Any

Aquesta taula s'omple manualment. Per tant el procés consta només de dos passos, el de càrrega de dades directa a través del Data Grid i una sortida cap a la base de dades.

Figura 22. Esquema general ETL any

4.2.2.2. Mes

La taula Mes també s'omple manualment. Per tant el procés consta només de dos passos, el de càrrega de dades directa a través del Data Grid i una sortida cap a la base de dades.

Figura 23. Esquema general ETL Mes

4.2.2.3. Estació

Igual que en els apartats anteriors, aquesta taula s'omple manualment. Per tant el procés consta només de dos passos, el de càrrega de dades directa a través del Data Grid i una sortida cap a la base de dades.

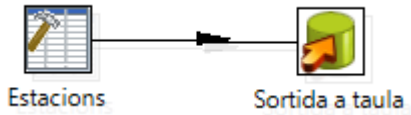


Figura 24. Esquema general ETL Estació

4.2.2.4. Establiment

Es tracta de la última taula omplerta de manera manual. El procés consta només de dos passos, el de càrrega de dades directa a través del Data Grid i una sortida cap a la base de dades.

Figura 25. Esquema general ETL Establiment

4.2.2.5. Municipi

El primer pas és extreure els noms dels municipis de Catalunya. Aquests noms s'extreuen del fitxer Infraestructura turística.xls.

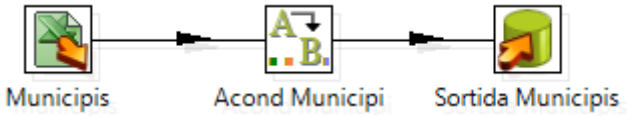


Figura 26. Primer esquema ETL Municipi

Els noms estan en la columna A, a partir del registre 8 i fins el 954 inclòs. Observem que algunes poblacions porten un article (primera lletra en majúscula) que està especificat després del nom de la població i una coma. Ja que aquetes dades estan extretes d'un organisme català oficial es donaran per bones i s'introduiran tal qual a la base de dades, però canviant l'article en minúscula per la majúscula.

El següent pas consisteix en afegir-hi informació sobre els codis postals, comarca, marca i comunitat. Llegim tota aquesta informació del fitxer Equipaments.csv. Netegem els camps del municipi, codi postal i eliminem les columnes innecessàries. Com que un municipi pot tenir diversos codis postals, primer cal eliminar aquells que estan buits i després utilitzar només un dels existents pel municipi. Finalment cal afegir la columna comunitat i actualitzar la base de dades.

La recepta queda de la següent manera:

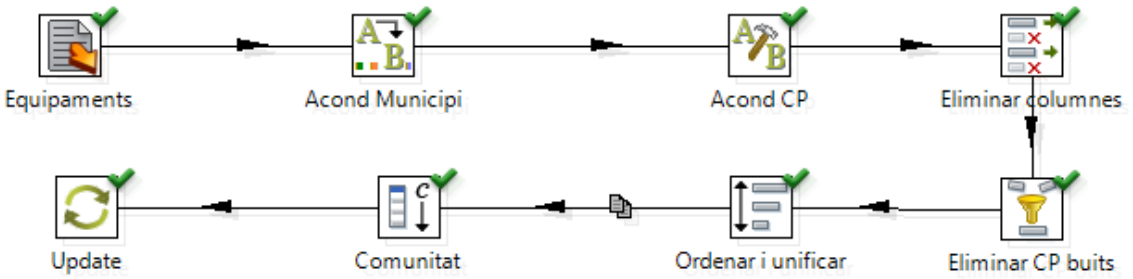


Figura 27. Segon esquema ETL Municipi

Una vegada executat el procediment anterior s'ha vist que s'han realitzat 935 actualitzacions, però a la nostra base de dades de municipis n'hi ha 947, per tant no tots els registres han estat actualitzats. Per això utilitzem un procediment per entrar els 12 municipis restants. Aquet procediment consisteix en seleccionar els municipis que no tenen dades, assignar-les i actualitzar la base de dades.

La recepta d'actualització dels camps restants és la següent:

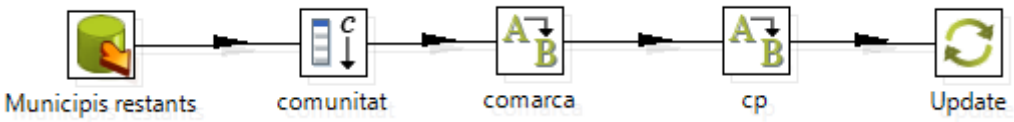


Figura 28. Tercer esquema ETL Municipi

Finalment es procedeix a la creació de la última transformació. Cal llegir la taula municipis, entrar manualment la relació de marques – comarques, fer un producte cartesià d'aquests, descartar els camps innecessaris i actualitzar la base de dades.

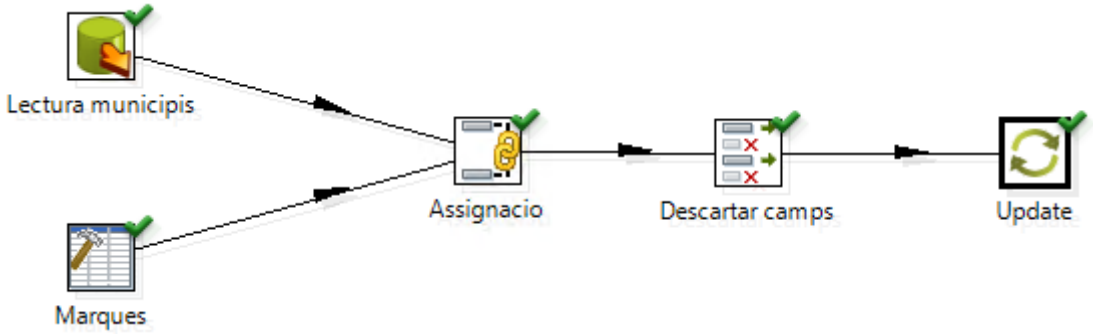


Figura 29. Quart esquema ETL Municipi

Amb aquest pas la taula es dóna per completa i finalitza el seu procés ETL.

4.2.2.6. Equipament

En aquest pas relacionarem els diferents equipaments amb el municipi al que pertanyen. No ens interessa saber els detalls dels equipaments, coneixent la categoria i sub-categoria en tenim prou. Si hi ha registres repetits els sumaritzarem, guardant la suma en una nova columna.

La recepta consisteix en llegir el fitxer Equipaments.xls, eliminar columnes innecessàries i adaptar el nom dels municipis segons l'experiència dels apartats anteriors. Seguidament els equipaments es divideixen en diferents camps, les seves paraules es separaran amb un espai en dos categories superiors i s'elimina la columna que conté només la paraula "Equipament". Després es fa la suma dels registres duplicats i es guarda en una nova columna. Així ja tenim la informació de taula que volem.

Finalment només cal creuar les dades dels municipis amb els seus respectius identificadors, eliminar la columna municipi i gravar els resultats en la taula Equipament.

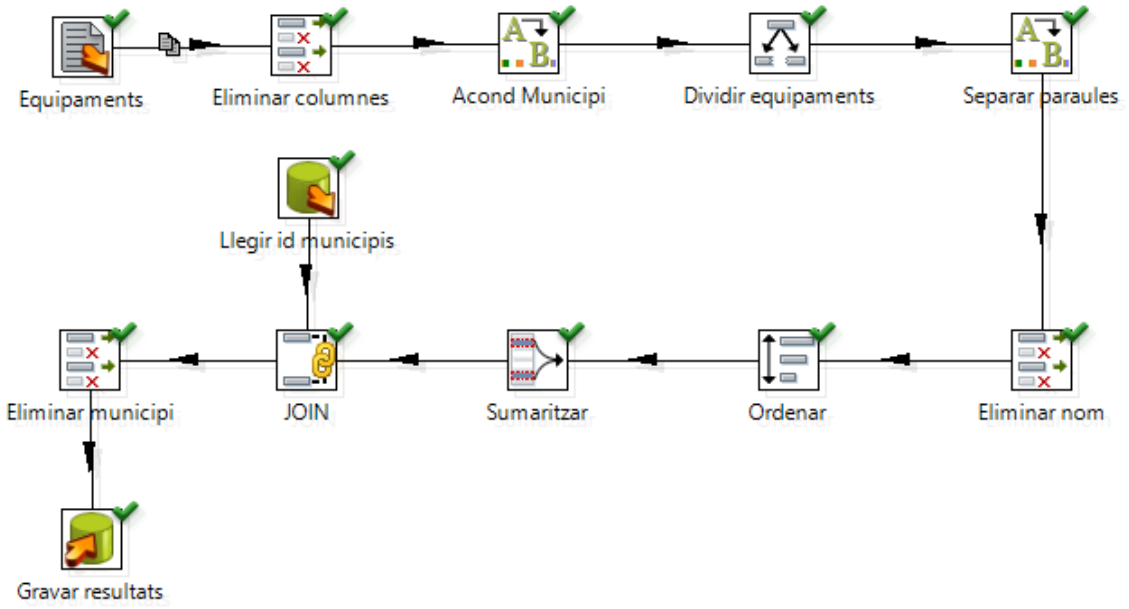


Figura 30. Esquema general ETL Equipament

4.2.2.7. Pernoctacions

La taula pernoctacions representa un Fet i és molt extensa ja que guarda dues mesures diferents – el número de pernoctacions i el número de viatgers. Per aquesta raó el procés ETL es divideix en dues parts, cadascuna referent a la seva mesura corresponent.

Comencem per pernoctacions. Abans de res cal destacar que treballem en un termini d'un any, per tant que l'esquema elaborat haurà de ser executat i adaptat per cada any, tot i que aquí es mostra el procés corresponent només a l'any 2011.

L'esquema de la recepta es pot dividir en dues parts. La primera és l'encarregada d'extreure dades de les places hoteleres i calcular la relació de places en un municipi respecte el total de la marca turística.

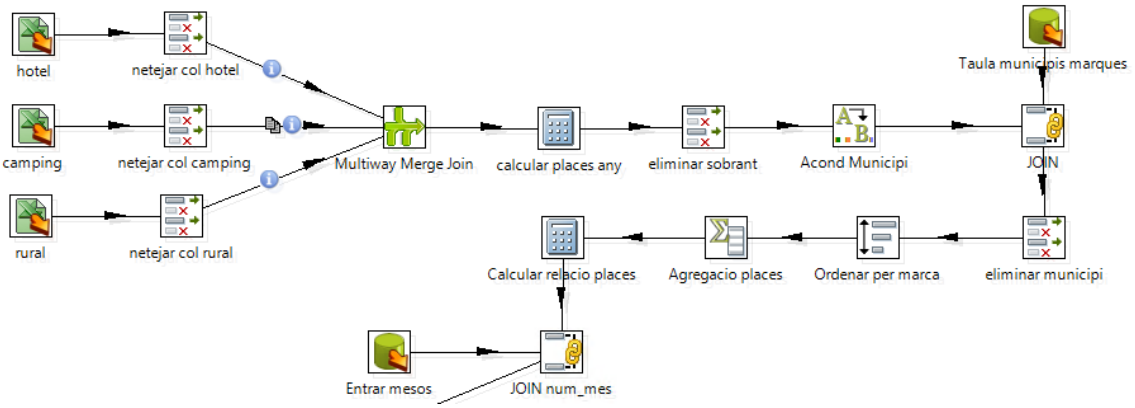


Figura 31. Primera part d'esquema ETL Pernoctacions

Tant els hotels, com els càmpings i els allotjaments rurals s'extreuen del mateix fitxer – Infraestructura turística.xls. Ara bé, cada tipus apunta als seus fulls i camps respectius. Després descarten els camps innecessaris.

Totes tres fonts s'uneixen en una sola, es realitza el càlcul total de places per any, s'eliminen les comunes sobrants i s'adapten els noms de municipis segons les experiències anteriors.

Seguidament s'introdueixen les marques, es relacionen amb els municipis, s'adapten les dades, es realitza el càlcul de places en una marca i la relació de places en un municipi respecte el total de la marca. Finalment les dades es creuen amb els diferents mesos copiant les dades entre aquests.

La segona part de l'esquema s'encarrega de recollir les dades relacionades amb les pernoctacions, creuar-les amb les anteriors, repartir les pernoctacions entre els municipis segons les places turístiques i acabar-les de preparar per entrar a la base de dades.

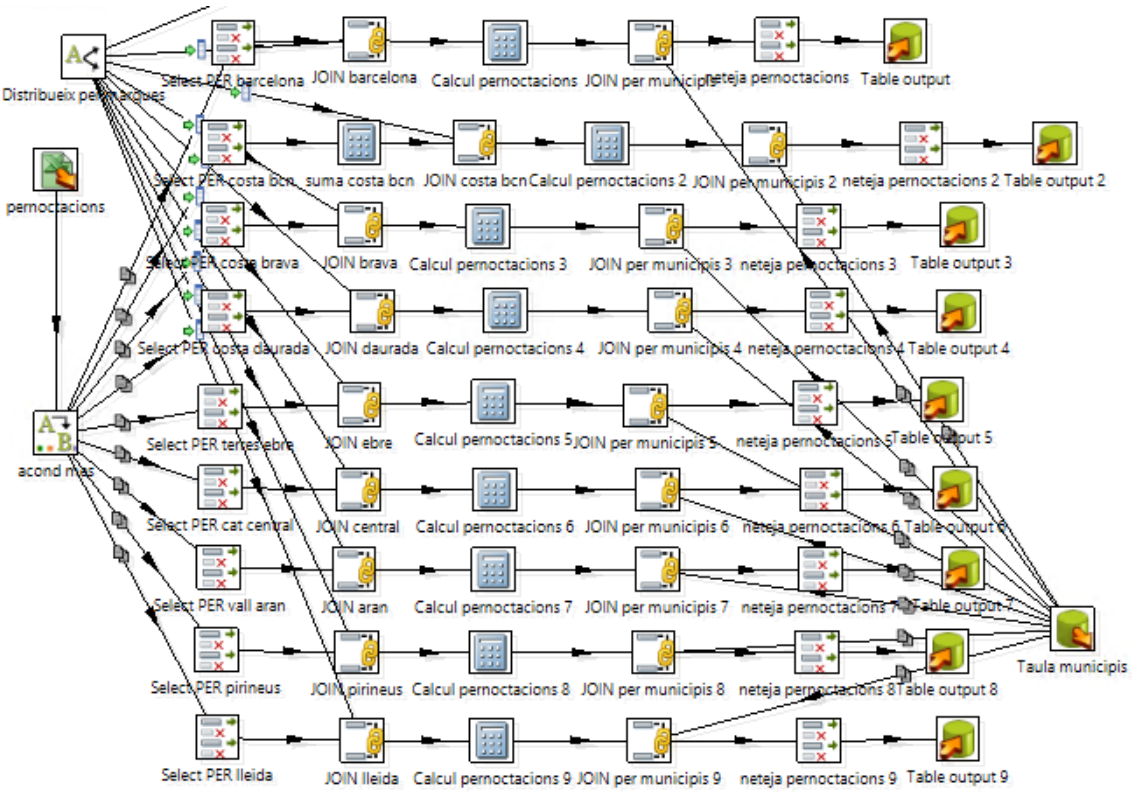


Figura 32. Segona part d'esquema ETL Pernoctacions

Primer les dades del pas anterior es distribueixen entre diferents marques turístiques. Després s'extreuen i s'adapten les dades de pernoctacions. Per comoditat de treball, en el mateix document d'Excel s'han copiat les dades de pernoctacions i s'han enganxat en un nou full transposades.

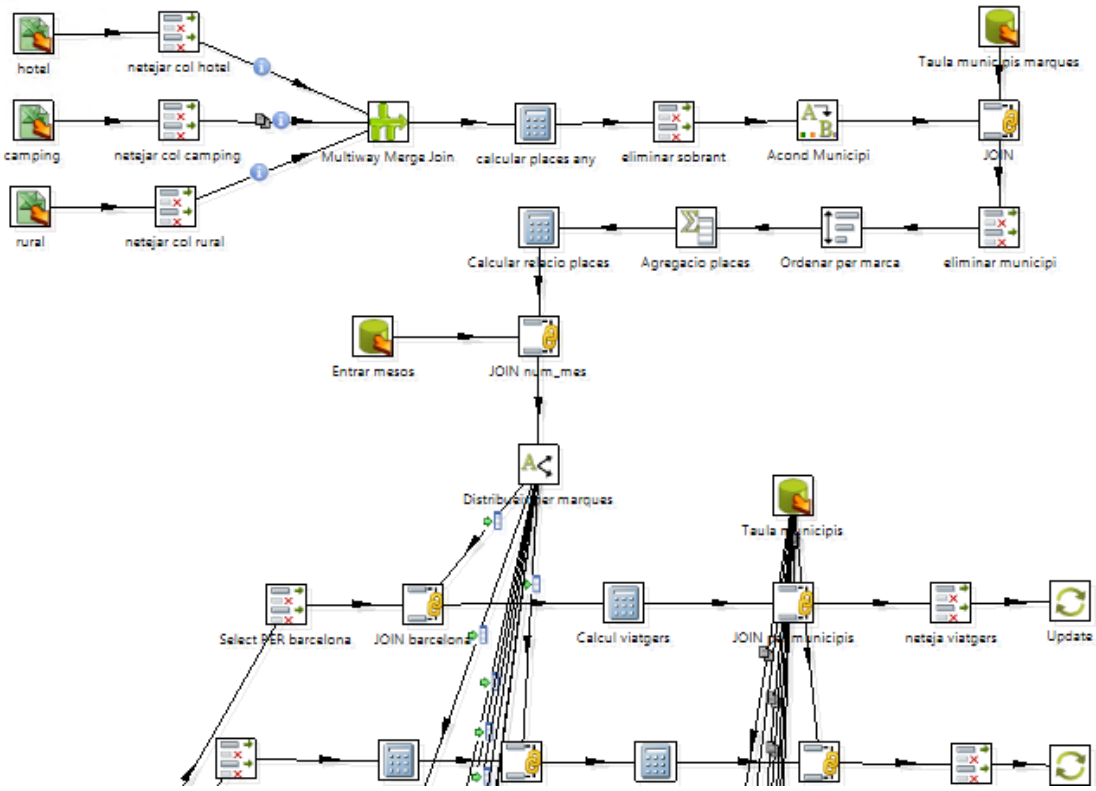
El següent pas és seleccionar només els valors corresponents a cada marca i creuar-los amb la sortida de la primera part, fer el càlcul de pernoctacions per municipi segons la relació de places calculada anteriorment i preparar les dades per la seva inserció dins de la base de dades, o sigui, creuar-les amb els identificadors de municipis, treure columnes innecessàries i entrar-ho tot dins de la base de dades.

Cal realitzar exactament el mateix procediment per totes les marques, menys per la Costa del Garraf i Costa-Maresme. En aquest cas les dues marques s'uneixen en una sola – Costa Barcelona.

De la mateixa manera s'han fet els anys posteriors, el 2012 i 2013, però sense calcular la suma de les dues marques (Costa del Garraf i Costa Barcelona-Maresme) en una sola Costa Barcelona.

La segona part del procés ETL de la mateixa taula consisteix en fer l'extracció, transformació, càlcul i càrrega del número de viatges. Es pot observar un paral·lelisme directe amb el cas anterior de les pernoctacions i per aquest motiu s'ha decidit utilitzar el mateix esquema, canviant els fitxers de pernoctacions pel de viatgers i actualitzant les files creades anteriorment.

La recepta d'un dels anys queda de la següent manera:



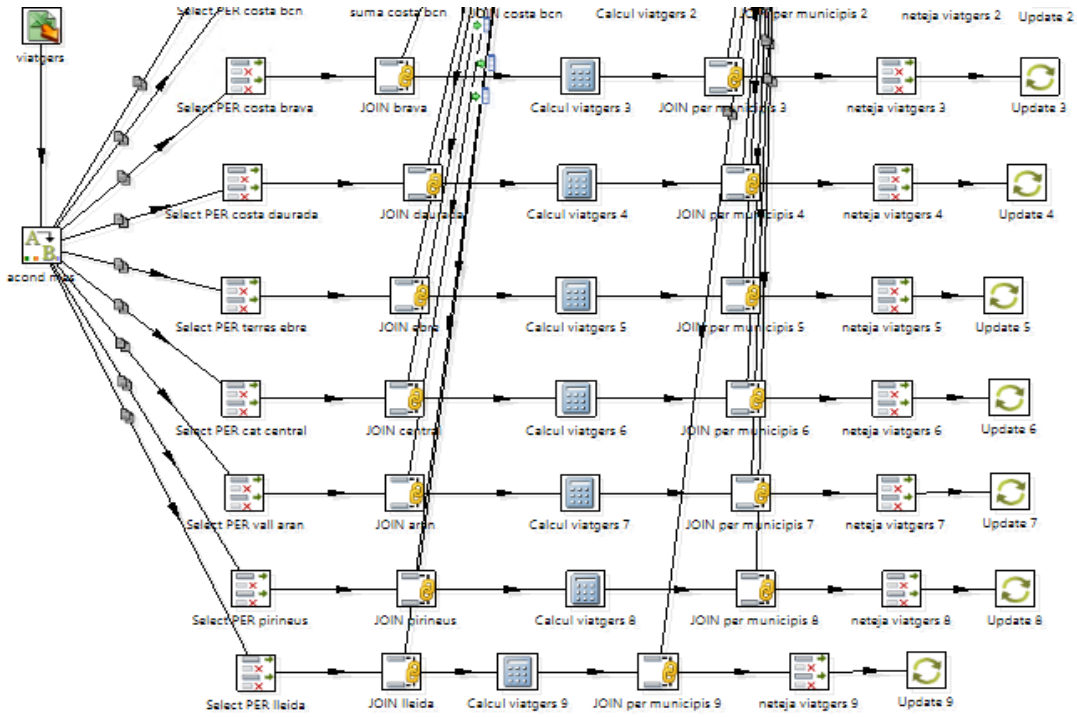


Figura 33. Esquema general ETL Viatgers

Podem observar que les úniques diferències amb l'esquema anterior són l'entrada de dades de pernoctacions per viatgers i actualització de la taula al final del procés en comptes d'insercions.

Igual que en el cas del càlcul de les pernoctacions cal repetir l'esquema amb els anys 2012 i 2013.

4.2.2.8. Infraestructures

Al tractar-se d'una taula Fet complexa amb moltes mesures es farà una introducció de dades progressiva.

Primer s'introdueixen tots els registres que apareixen de manera granular a més baix nivell. Aquest és el cas del conjunt de camps: any_, municipi, establiment, numEstabliments i placesEstabliments. La introducció es fa de manera reiterativa per anys. Introduïm primer les dades de any_, municipi, establiment i numEstabliments.

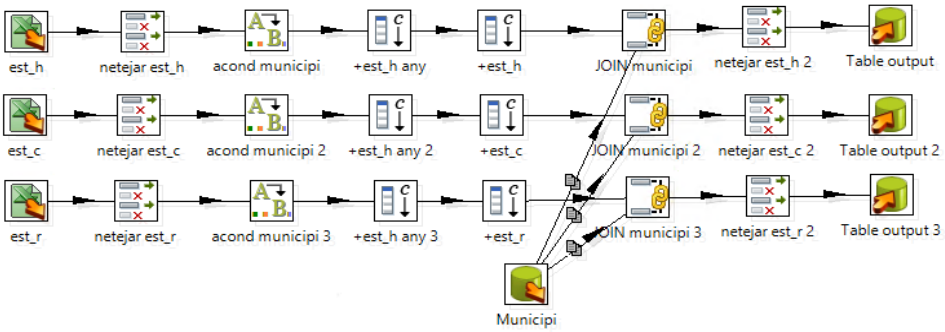


Figura 34. Primer esquema ETL infraestructures

Al principi extraiem les dades dels establiments de cada any i municipi. Eliminem els camps sobrants i adaptem els noms de municipis. Seguidament afegim l'any del que es tracta i una columna amb l'identificador de l'establiment: 1 per hotels, 2 per càmpings i 3 per turisme rural. Finalment només queda creuar les dades de municipis amb els seus respectius identificadors, adaptar les taules i introduir-les dins de la base de dades. Es pot observar que el mateix procediment es repeteix per cada tipus d'establiment.

Després es procedeix a calcular i actualitzar el camp del número de places per any, municipi i tipus d'establiment. Per això s'utilitza el mateix esquema que en el pas anterior, menys per la sortida, on en comptes d'una inserció es realitza una actualització.

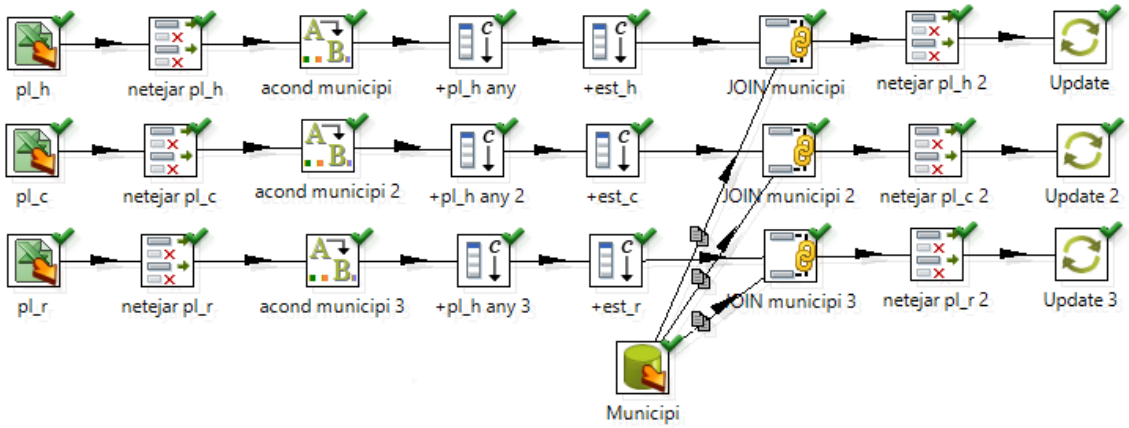


Figura 35. Segon esquema ETL infraestructures

El pròxim pas consisteix en introduir el número de policies. Aquesta dada s'extreu directament del fitxer Policies locals.xls per cada any treballat. S'eliminen les columnes innecessàries, els noms de municipis s'homogeneïtzen i tot s'entra a la base de dades a través de les actualitzacions. La recepta és la següent:

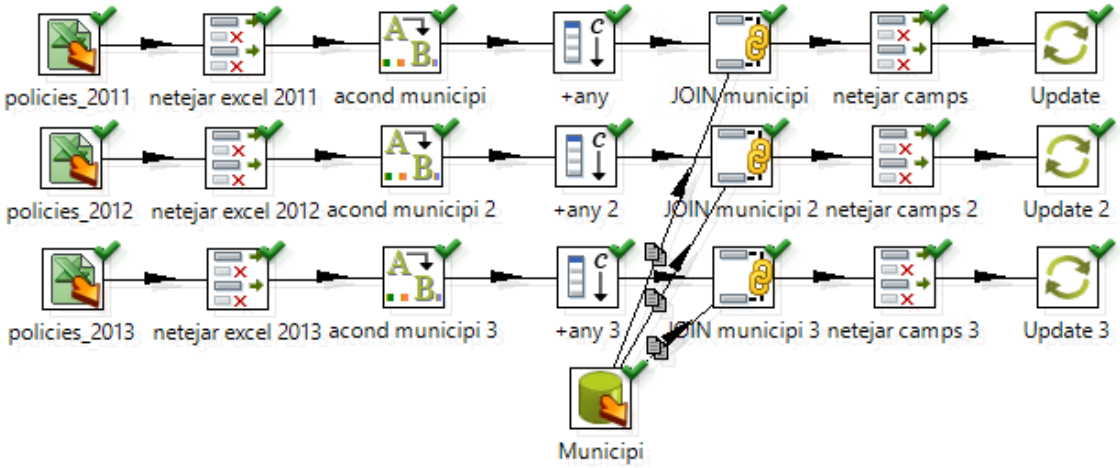


Figura 36. Tercer esquema ETL infraestructures

Es veu clarament que la recepta és molt semblant a l'anterior, canviant les dades de partida.

L'últim pas és introduir el cens de la població en cada municipi per any. Aquestes dades es troben en els fitxers pobmun1X.xls. Com que aquests fitxers són molt grans, extraurem en un fitxer extern (pob1X.xls) només les dades corresponents a la població total a Catalunya, o sigui províncies de Girona, Tarragona, Lleida i Barcelona.

	A	B	C	D	E	F	G
1	**Cifras de población resultantes de la Revisión del Padrón municipal a 1 de enero**						
2	**CPRO**	**PROVINCIA**	**CMUN**	**NOMBRE**	**AMBOS SEXOS**	**VARONES**	**MUJERES**
3	08	Barcelona	001	Abrera	11.611	5.926	5.685
4	08	Barcelona	002	Aguilar de Segarra	245	131	114
5	08	Barcelona	003	Alella	9.570	4.705	4.865
6	08	Barcelona	004	Alpens	302	165	137
7	08	Barcelona	005	Ametlla del Vallès, L'	8.111	4.081	4.030
8	08	Barcelona	006	Arenys de Mar	14.863	7.307	7.556

Figura 37. Retall d'Excel pob1X.xls

La recepta és la mateixa per tots els anys. Primer s'extreuen els camps del fitxer, s'adapta el nom del municipi, s'afegeix l'any, les dades es creuen amb l'identificador de la taula municipi, es seleccionen els camps i s'actualitza la base de dades.

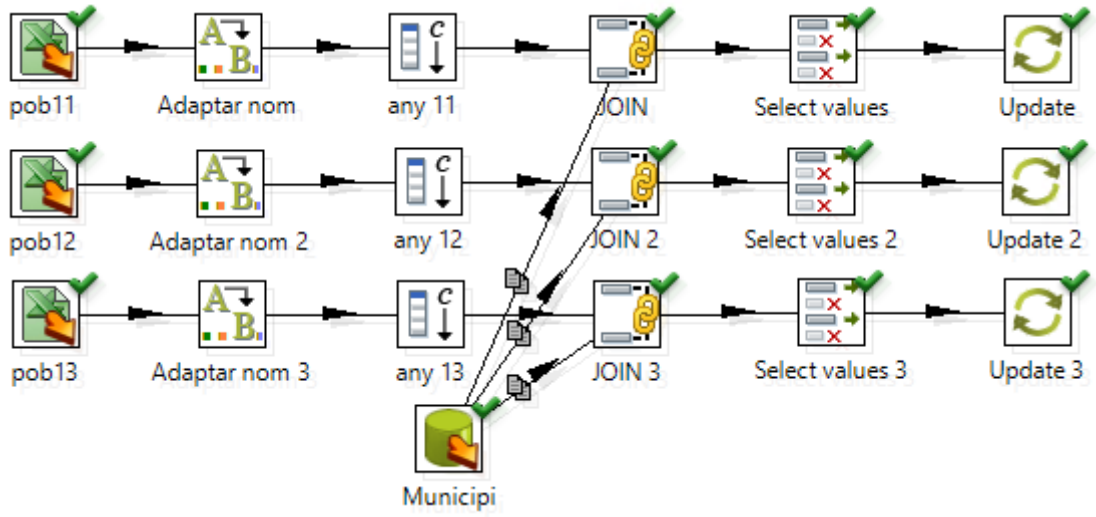


Figura 38. Quart esquema ETL infraestructures

4.3. INFORMES

4.3.1. ENTORN

L'entorn de treball és el Pentaho Report Designer 5.1.

Per defecte aquest entorn no incorpora el controlador de connexió amb el MySQL. Per això s'ha de copiar manualment el fitxer **mysql-connector-java-5.1.26.jar** dins de la carpeta **C:\Program Files\report-designer\lib**.

En cada nou informe que definim cal crear una nova connexió amb els següents paràmetres:

Paràmetre	Valor
Connection Name:	DW
Connection Type:	MySQL
Access:	Native (JDBC)
Host Name:	localhost
Database Name:	tfc_dw
Port Number:	3306
User Name:	normal
Password:	TFC2014

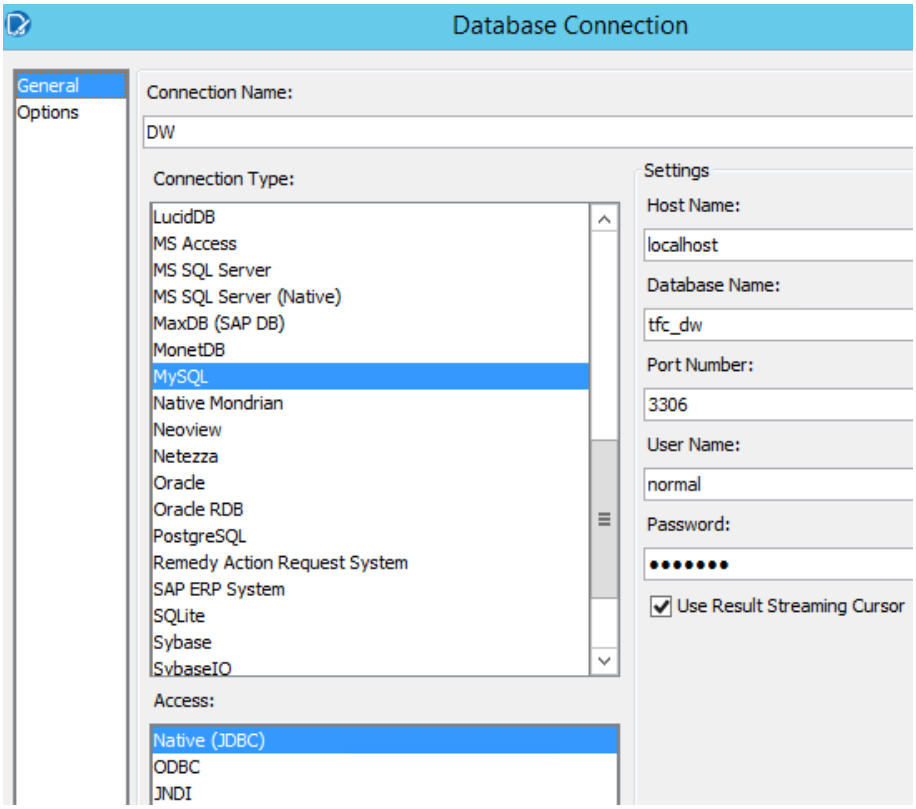


Figura 39. Connexió del Report Designer amb la base de dades

4.3.2. CREACIÓ D'INFORMES

A continuació s'exposaran les característiques més destacades en la creació dels informes demanats.

4.3.2.1. Rànquing de municipis per categoria d'equipaments

L'informe comença amb un gràfic que mostra el top 5 de municipis per número d'equipaments d'una categoria determinada. Seguidament es mostra el rànquing detallat de municipis per categoria d'equipaments, de major a menor.

El gràfic s'implementa amb un sub-informe a partir de la consulta:

```
SELECT m.municipi, sum(e.numero) as 'total'
FROM municipi m, equipament e, (SELECT @curRank := 0) r
WHERE m.id=e.municipi
      AND e.equipament=${eq}
GROUP BY m.municipi
ORDER BY total DESC
LIMIT 5
```

El detall del rànquing s'extreu de la consulta:

```
SELECT @curRank := @curRank + 1 as 'rank', municipi, total
FROM
(SELECT m.municipi, sum(e.numero) as 'total'
FROM municipi m, equipament e, (SELECT @curRank := 0) r
WHERE m.id=e.municipi
      AND e.equipament=${eq}
GROUP BY m.municipi
ORDER BY total DESC) as consulta
```

En executar l'informe a l'usuari se li demana que seleccioni un equipament a partir d'una llista predefinida. La consulta per extreure la llista:

```
SELECT equipament
FROM equipament
GROUP BY equipament
ORDER BY equipament ASC
```

El paràmetre seleccionat es guarda com a eq i és passat al sub-informe del gràfic.

4.3.2.2. Màxim i mínim d'efectius policials per tipologia d'establiment i municipi

L'informe es divideix en dues parts. La primera representa un gràfic amb el seu corresponent detall dels municipis amb màxim d'efectius policials per cada tipus d'establiment a partir del rati calculat (policies / establiments). La segona part és similar, però en aquest cas representant el mínim (deixant de banda els municipis on el rati és 0 o no es pot calcular).

La consulta encarregada del primer sub-informe:

```
SELECT m.municipi, i.numPolicies, i.placesEstabliments,
      (i.numPolicies/i.placesEstabliments) as 'rati'
FROM infraestructures i, municipi m
```

```
WHERE i.municipi=m.id
      AND i.any_=${any_}
      AND i.establiment=${id_establ}
      AND (i.numPolicies/i.placesEstabliments) IS NOT NULL
      AND (i.numPolicies/i.placesEstabliments) <> 0
GROUP BY municipi
ORDER BY rati DESC
LIMIT 5
```

La consulta encarregada del segon sub-informe:

```
SELECT m.municipi, i.numPolicies, i.placesEstabliments,
       (i.numPolicies/i.placesEstabliments) as 'rati'
FROM infraestructures i, municipi m
WHERE i.municipi=m.id
      AND i.any_=${any_}
      AND i.establiment=${id_establ}
      AND (i.numPolicies/i.placesEstabliments) IS NOT NULL
      AND (i.numPolicies/i.placesEstabliments) <> 0
GROUP BY municipi
ORDER BY rati ASC
LIMIT 5
```

En executar l'informe a l'usuari se li demanen dos paràmetres, l'any de la consulta i el tipus d'establiment del que es tracta. Les consultes per obtenir aquests valors:

```
SELECT any_
FROM any_
ORDER BY anv  ASC
```

```
SELECT id, establiment
FROM establiment
```

4.3.2.3. Rati de policies locals per habitant

En aquest informe es mostren al detall tots els municipis de Catalunya, segons l'any seleccionat per l'usuari, amb el seu cens d'habitants, policies i el càlcul d'habitants que pertoquen per cada policia en el municipi.

Tots els registres es treuen a partir de la següent consulta:

```
SELECT m.municipi, i.numPolicies, i.cens,
      (i.cens/i.numPolicies) as 'rati'
```

```
FROM infraestructures i, municipi m
WHERE i.municipi=m.id
      AND i.any_=${any_}
GROUP BY municipi
ORDER BY municipi ASC
```

Pel que fa al paràmetre any, igual que en els informes anteriors es treu de la consulta:

```
SELECT any_
FROM any_
ORDER BY any_ ASC
```

Cal destacar que en molts municipis el promig d'habitants per policia és molt alt, mentre altres directament no disposen de cap policia. Això és degut a que molts municipis petits no disposen de policia local pròpia i se'ls serveix policia d'altres municipis.

Per aquesta raó es podria augmentar la zona geogràfica, per exemple a comarca, on els policies es repartirien entre tots els habitants de la mateixa. En aquest la consulta seria de la següent manera:

```
SELECT m.municipi, c.rati
FROM municipi m,
      (SELECT m.comarca, sum(i.numPolicies/3) as 'policies',
      sum(i.cens/3) as 'cens',
      CAST(sum(i.cens/3)/sum(i.numPolicies/3) as UNSIGNED) as
      'rati'
      FROM infraestructures i, municipi m
      WHERE i.municipi=m.id
            AND i.any_=${any_}
      GROUP BY m.comarca
      ORDER BY rati) as c
WHERE m.comarca=c.comarca
```

En aquest punt s'ha detectat el problema que algunes comarques no tenen ni un sol policia assignat, o d'altres que en tenen sospitosament pocs. Per aquesta raó caldrà considerar les dades dels policies com a incompletes.

4.3.2.4. Distribució mensual i estacionalitat de les pernoctacions

Per visualitzar les dades demanades s'ha preparat un informe que resumeix les dades de cada marca turística al llarg dels tres anys. L'informe s'ha estructurat en grups de diferents marques.

En la primera part de manera visual es mostra un gràfic de barres amb les dades dels tres anys d'estudi repartides en 12 mesos que té un any. Aquesta distribució permet detectar d'una manera molt senzilla i visual les diferents estacionalitats existents. La consulta a partir de la qual es dibuixa la gràfica és la següent:

```
SELECT m.marca, mes.any_, mes.nomMes, sum(p.numPernoctacions) as
      'pern'
FROM pernoctacions p, municipi m, mes
WHERE p.municipi=m.id AND p.mes=mes.id
GROUP BY m.marca, p.mes
```

La segona part de l'informe, a partir d'un sub-informe, implementa una taula resumida amb les pernoctacions dels tres anys dividits en els 12 mesos de l'any. Aquesta segona part serveix com un complement numèric més exacte que la primera part. La consulta a partir de la qual es construeix la taula és la següent:

```
SELECT  c1.nomMes as 'mes', c1.pern2011, c2.pern2012, c3.pern2013
FROM
      (SELECT mes.nomMes, sum(p.numPernoctacions) as 'pern2011'
      FROM pernoctacions p, municipi m, mes
      WHERE p.municipi=m.id AND p.mes=mes.id
            AND m.marca=${marca}
            AND mes.any_=2011
      GROUP BY p.mes) as c1,
      (SELECT mes.nomMes, sum(p.numPernoctacions) as 'pern2012'
      FROM pernoctacions p, municipi m, mes
      WHERE p.municipi=m.id AND p.mes=mes.id
            AND m.marca=${marca}
            AND mes.any_=2012
      GROUP BY p.mes) as c2,
      (SELECT mes.nomMes, sum(p.numPernoctacions) as 'pern2013'
      FROM pernoctacions p, municipi m, mes
      WHERE p.municipi=m.id AND p.mes=mes.id
            AND m.marca=${marca}
            AND mes.any_=2013
      GROUP BY p.mes) as c3
WHERE c1.nomMes=c2.nomMes AND c2.nomMes=c3.nomMes
```

4.3.2.5. % de pernoctacions d'un municipi sobre el total

Aquest informe té com a objectiu mostrar el percentatge de pernoctacions d'un municipi sobre la resta de municipis, en un temps determinat. Per tenir una major perspectiva s'ha definit la granularitat del temps a un any.

La primera part de l'informe és un sub-informe encarregat de mostrar de manera visual la distribució dels percentatges a partir de la següent consulta:

```
SELECT m.municipi, sum(p.numPernoctacions) as 'pern'
FROM pernoctacions p, municipi m, mes
WHERE p.municipi=m.id AND p.mes=mes.id
      AND mes.any_=${any_}
      AND m.id=${municipi_}
GROUP BY m.municipi
UNION
SELECT 'Resta', sum(p.numPernoctacions)-pern
FROM pernoctacions p,
      (SELECT m.municipi, sum(p.numPernoctacions) as 'pern'
      FROM pernoctacions p, municipi m, mes
      WHERE p.municipi=m.id AND p.mes=mes.id
            AND mes.any_=${any_}
            AND m.id=${municipi_}
      GROUP BY m.municipi) as c
```

Seguidament hi ha un detall de les dades, tret a partir de la consulta:

```
SELECT sum(p.numPernoctacions) as 'pern', c.total,
CAST(sum(p.numPernoctacions)/c.total * 100 as DECIMAL(4,2)) as '%'
FROM pernoctacions p, municipi m, mes,
(SELECT sum(numPernoctacions) as 'total' FROM pernoctacions) as c
WHERE p.municipi=m.id AND p.mes=mes.id
      AND mes.any_=${any_}
      AND m.id=${municipi_}
GROUP BY m.municipi
```

El paràmetres que capten l'any i el municipi es recullen amb les seves consultes corresponents:

```
SELECT any_
FROM any_
ORDER BY any_ ASC
```

```
SELECT id, municipi
FROM municipi
```

4.3.2.6. "Top ten" de municipis per franja de pernoctacions

Aquest informe classifica tots els municipis segons el número de pernoctacions en mesos determinats, que al seu torn s'agrupen en estacions. La temporalitat

escollida és de tots els anys dels quals es disposa de dades (tres anys). La granularitat de la zona és de municipi.

La consulta principal és la següent:

```
SELECT @curRank := @curRank + 1 as 'rank', municipi, pern
FROM
      (SELECT m.municipi, sum(numPernoctacions) as 'pern'
      FROM pernoctacions p, municipi m, mes, estacio e,
            (SELECT @curRank := 0) r
      WHERE p.municipi=m.id AND p.mes=mes.id AND e.mes=mes.id
            AND e.estacio IN (${estacio})
      GROUP BY m.id
      ORDER BY pern DESC
      LIMIT 10) as c
```

En executar l'informe a l'usuari se li demana que seleccioni una o diverses estacions, que s'extreuen a partir de la consulta:

```
SELECT estacio
FROM estacio
GROUP BY estacio ASC
```

4.3.2.7. Total de pernoctacions estimades per municipi

L'informe es basa en una sola consulta que calcula l'estimació de pernoctacions per tots els municipis catalans per l'any 2014:

```
SELECT c.id, m.municipi, ROUND(sum(parcials)) as 'pern2014'
FROM municipi m,
      (SELECT p.municipi as 'id',
            CASE any_
                  WHEN 2011 THEN sum(p.numPernoctacions)*0.2
                  WHEN 2012 THEN sum(p.numPernoctacions)*0.3
                  WHEN 2013 THEN sum(p.numPernoctacions)*0.5 END as
                  'parcials'
      FROM pernoctacions p, mes
      WHERE p.mes=mes.id
      GROUP BY p.municipi, mes.any_) as c
WHERE m.id=c.id
GROUP BY c.id
```

4.3.2.8. Promig de viatgers per tipus d'establiment i comarca

Aquest informe suposa un repte a nivell de distribució de dades. Cal repartir els diferents viatgers que arriben a un municipi per tipus d'establiments segons la

proporció del número de places d'aquests establiments en el municipi. Després cal generalitzar els resultats a nivell de la comarca. L'usuari ha d'indicar quin tipus d'establiment li interessa i de quin any vol veure els resultats. L'informe consisteix en un llistat de totes les comarques amb el seu respectiu promig de viatgers.

Només hi ha una consulta general que inclou diverses sub-consultes:

```
SELECT m.comarca, sum(viatgersPerEstabliment) as 'promigViatgers'
FROM (SELECT c2.id, c3.establiment, CAST(c2.viatgers *
       (c3.placesEstabliments / c2.placesMunicipi) as DECIMAL(10,2))
       as 'viatgersPerEstabliment'
   FROM
      (SELECT id, viatgers, sum(placesEstabliments) as 'placesMunicipi'
      FROM (SELECT m.id, i.establiment, i.placesEstabliments,
            sum(p.numViatgers) as 'viatgers'
            FROM pernoctacions p, municipi m, infraestructures i, mes
            WHERE p.municipi=m.id AND i.municipi=m.id
                  AND p.mes=mes.id AND i.any_=mes.any_
                  AND i.any_=${any_}
            GROUP BY m.id, i.establiment) as c1
      GROUP BY id) as c2,
      (SELECT m.id, i.establiment, i.placesEstabliments
      FROM pernoctacions p, municipi m, infraestructures i, mes
      WHERE p.municipi=m.id AND i.municipi=m.id
            AND p.mes=mes.id AND i.any_=mes.any_
            AND i.any_=${any_}
      GROUP BY m.id, i.establiment) as c3
   WHERE c2.id=c3.id) as c4, municipi m
WHERE c4.id=m.id AND establiment=${establiment}
GROUP BY comarca
```

Els paràmetre establiment i any es treuen a partir de les seves respectives consultes:

```
SELECT id, establiment
FROM establiment
```

```
SELECT any_
FROM any_
ORDER BY anv  ASC
```

4.3.2.9. % d'ocupació per marca turística

En aquest informe es mostren totes les marques turístiques amb els seus respectius percentatges d'ocupació, dins de la temporalitat d'un mes concret.

El càlcul del percentatge es basa en les dades de pernoctacions i viatgers. Es calcula l'estància mitja, que l'INE[9] defineix com el coeficient entre el número de pernoctacions i el número de viatgers entrats, es multiplica pel número de viatgers i es divideix entre els dies del mes, fet que dóna el número de places ocupades. El percentatge d'ocupació és la relació entre el número de places ocupades i les disponibles. Tot plegat es resumeix en la següent consulta:

```
SELECT c1.marca, ROUND(placesOcupades/placesDisponibles * 100, 2)
      as 'ocupació'
FROM (SELECT m.marca,
      sum(numPernoctacions)/sum(numViatgers) as 'estancia',
      sum(numViatgers)*(sum(numPernoctacions)/sum(numViatgers)) /
      30 as 'placesOcupades'
      FROM pernoctacions p, mes, municipi m
      WHERE p.mes=mes.id AND p.municipi=m.id
            AND mes.id=${mes}
      GROUP BY m.marca) as c1,
       (SELECT m.marca,
            sum(placesEstabliments) as 'placesDisponibles'
      FROM infraestructures i, municipi m
      WHERE i.municipi=m.id
            AND any_=${any_}
      GROUP BY m.marca) as c2
WHERE c1.marca=c2.marca
```

Els paràmetres d'any i del mes s'extreuen de les consultes dependents entre si:

```
SELECT any_
FROM any_
ORDER BY any_  ASC
```

```
SELECT id, nomMes
FROM mes
WHERE any_=${any_}
```

4.3.2.10. Categorització de municipis A/B/C

L'informe fa una categorització de municipis segons l'anàlisi:

[9] Instituto Nacional de Estadística

A. El 20% dels municipis que representen el 80% de les pernoctacions. Són els imprescindibles pel negoci, la seva base.
B. El 30% dels municipis que representen el 15% de les pernoctacions. Són els candidats a convertir-se en els importants.
C. El 50% dels municipis que representen el 5% de les pernoctacions restants. Són molts municipis que tenen molt poc moviment turístic i l'atenció als quals ha de ser molt reduïda.

L'objectiu de l'informe és seleccionar una categoria i presentar els municipis que inclou. Les dades s'extreuen a partir de la següent consulta:

```
SELECT *
FROM
(SELECT @rank := @rank + 1 as 'rank', c.municipi, c.pern,
c.pernTotal, ROUND(c.percent, 4) as 'percent',
      CASE
            WHEN m.cont/100*20 > @rank THEN 'A'
            WHEN m.cont/100*50 > @rank THEN 'B'
            ELSE 'C' END as 'categoria'
FROM
      (SELECT m.municipi, sum(numPernoctacions) as 'pern',
      c.pernTotal,
      sum(numPernoctacions)*100/pernTotal as 'percent'
      FROM pernoctacions p, mes, estacio e, municipi m,
            (SELECT sum(numPernoctacions) as 'pernTotal'
            FROM pernoctacions p, mes, estacio e
            WHERE p.mes=mes.id AND mes.id=e.mes
                  AND e.estacio='estiu'
            GROUP BY e.estacio) as c
      WHERE p.mes=mes.id AND mes.id=e.mes AND p.municipi=m.id
            AND e.estacio='estiu'
      GROUP BY m.id
      ORDER BY pern DESC) as c,
      (SELECT count(*) as 'cont' FROM municipi) as m,
      (SELECT @rank := 0) r) as c
WHERE categoria=${cat}
ORDER BY rank
```

Per visualitzar millor les proporcions, a la primera part es presenta un gràfic del que representen el total de les pernoctacions d'una categoria. Això s'aconsegueix amb un sub-informe que s'alimenta a partir de la consulta:

```
SELECT ${cat} as 'categoria',
      ROUND(@cat := sum(percent), 2) as 'percent'
FROM
```

```
(SELECT @rank := @rank + 1 as 'rank', c.municipi, c.pern,
      c.pernTotal, ROUND(c.percent, 4) as 'percent',
      CASE
            WHEN m.cont/100*20 > @rank THEN 'A'
            WHEN m.cont/100*50 > @rank THEN 'B'
            ELSE 'C' END as 'categoria'
FROM
      (SELECT m.municipi, sum(numPernoctacions) as 'pern',
      c.pernTotal,
      sum(numPernoctacions)*100/pernTotal as 'percent'
      FROM pernoctacions p, mes, estacio e, municipi m,
            (SELECT sum(numPernoctacions) as 'pernTotal'
            FROM pernoctacions p, mes, estacio e
            WHERE p.mes=mes.id AND mes.id=e.mes
                  AND e.estacio='estiu'
            GROUP BY e.estacio) as c
      WHERE p.mes=mes.id AND mes.id=e.mes AND p.municipi=m.id
            AND e.estacio='estiu'
      GROUP BY m.id
      ORDER BY pern DESC) as c,
      (SELECT count(*) as 'cont'
      FROM municipi) as m,
      (SELECT @rank := 0) r) as c
WHERE categoria=${cat}
GROUP BY categoria
UNION
SELECT 'Resta', ROUND(100-@cat, 2)
```

En la categorització s'ha agafat com a dada de referència el primer percentatge, calculat sobre el número total de municipis. En l'informe es visualitza que la llei de Pareto no compleix estrictament amb les nostres dades, ja que per exemple el 20% de municipis representa molt més que el 80% de pernoctacions.

4.4. LLANÇAMENT A PRODUCCIÓ

4.4.1. CONFIGURACIÓ DEL SERVIDOR

En el servidor cal configurar un nou usuari per l'accés als avaluadors del treball que només té permisos de lectura.

Primer s'ha accedit a la consola d'administració amb usuari **Admin** i contrasenya **password**. Una vegada a dins, cal desplaçar-se a l'apartat d'administració. Ens portarà a l'apartat d'usuaris, on s'ha d'indicar que volem afegir un nou usuari, el seu nom – **usuari** i contrasenya – **TFC2014**.

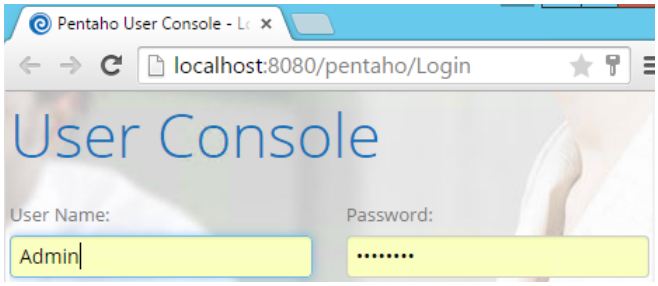


Figura 40. Pantalla d'accés de Pentaho

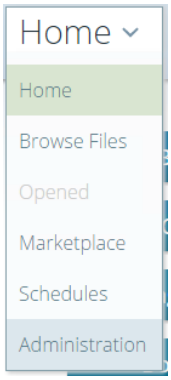


Figura 41. Apartat d'administració de Pentaho

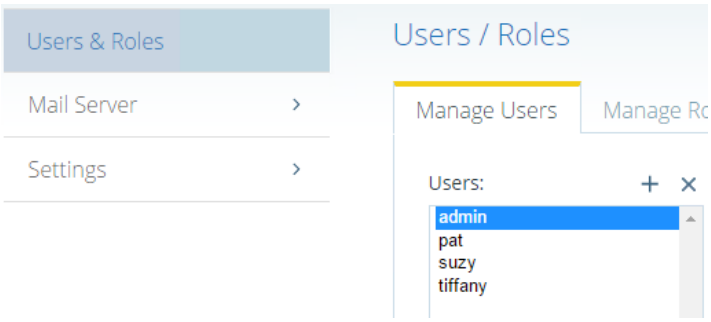


Figura 42. Gestió de usuaris de Pentaho

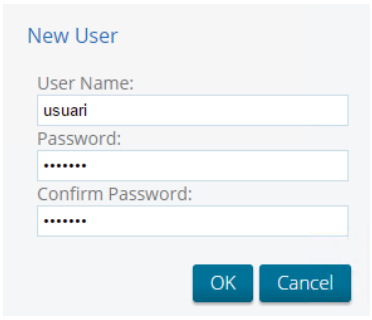


Figura 43. Afegir un nou usuari a Pentaho

El següent pas consisteix en assignar permisos al nou usuari. En la pestanya *Manage Roles* cal afegir un nou rol – *Auditor* (només permisos de lectura) i assignar-li aquest rol a l'usuari creat prèviament.

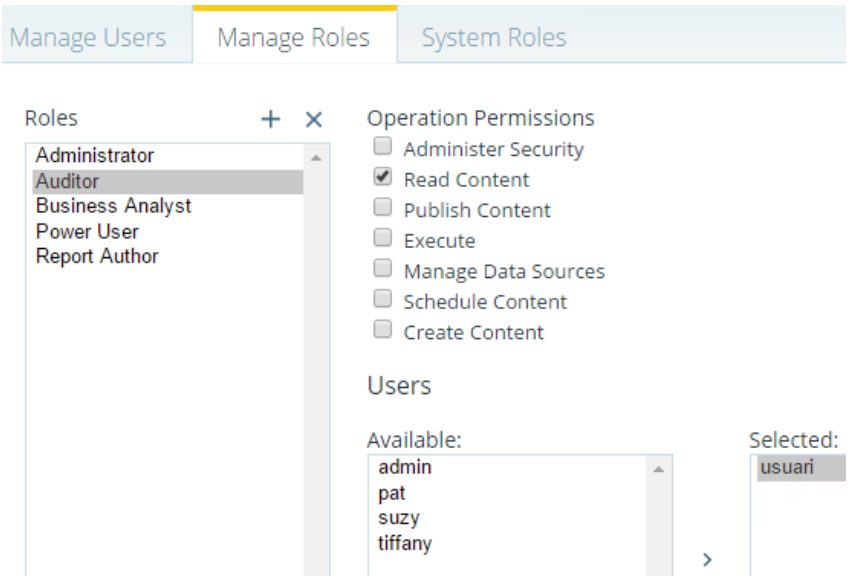


Figura 44. Gestió de rols en Pentaho

4.4.2. CÀRREGA DELS INFORMES

La càrrega es realitza directament des del propi Pentaho Report Designer. Una vegada tenim l'informe elaborat cal anar a *File->Publish...*, seleccionar les dades de la connexió amb el servidor, indicar les opcions de l'informe (important indicar correctament la carpeta de l'usuari que hi accedirà) i acceptar.

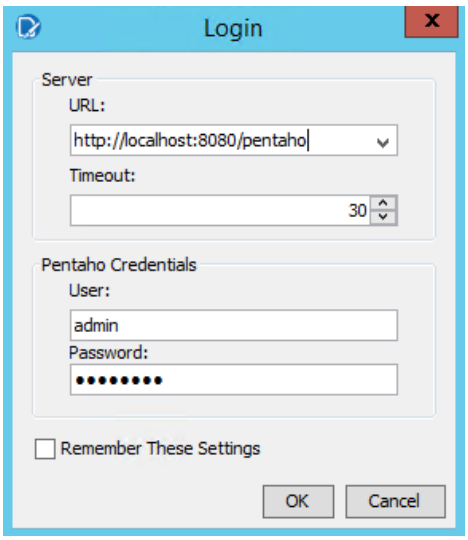


Figura 45. Accés al servidor des del Report Designer

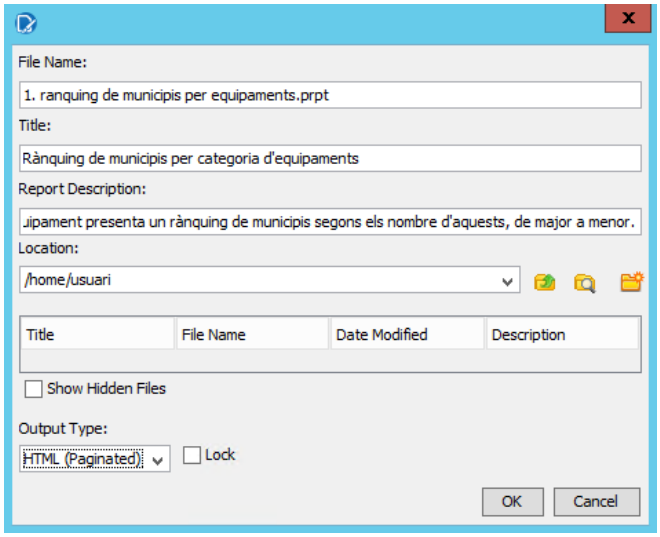


Figura 46. Configuració d'informe a pujar cap al servidor

Aquest procediment s'ha de repetir per tots els informes.

5. MILLORES I FUTURES LÍNIES DE TREBALL

En aquest projecte s'ha aconseguit una solució escalable que permet moltes millores i noves implementacions. Algunes de les més destacades es detallen a continuació.

- **Automatització de tasques ETL**. Els procediments ETL desenvolupats es podrien millorar per tenir més independència amb les dades entrants, per no haver d'aplicar pre-edicions manuals abans d'entrar en el cicle ETL. De fet, es podria contactar amb els organismes proveïdors de dades i acordar un enriquiment periòdic de les mateixes en un format estàndard i constant. Llavors només caldria crear JOBs en Pentaho Spoon perquè cada cert temps enriquís la base de dades. D'aquesta manera es podria tenir el DW contínuament actualitzat sense cap esforç de més.

- **Anàlisi OLAP**. El DW implementat en MySQL segueix totes les directrius necessàries per poder executar tot tipus d'anàlisi de dades. De moment s'han implementat només la part dels informes com a interfície d'interacció amb l'usuari. Si bé, alguns prefereixen la senzillesa dels informes pre-definits, d'altres voldrien disposar d'eines analítiques més dinàmiques. Per aquesta raó el següent pas lògic seria implementar anàlisi de cubs OLAP en el servidor Pentaho BI Server a través de Mondrian i Saiku.

- **Quadres de comandament**. Perquè la direcció pugui prendre decisions al moment ha de tenir la informació clau sempre disponible, de manera clara i senzilla. Per aquesta raó es poden implementar quadres de comandament via web, que siguin multiplataforma i disponibles a totes hores.

- **Integració amb altres sistemes BI empresarial**. La solució desenvolupada pot augmentar significativament la seva productivitat i eficiència si s'integra amb els altres sistemes BI disponibles a l'empresa, com poden ser sistemes de mineria de dades, CRM, GIS, gestió de coneixement, etc.

6. CONCLUSIONS

El projecte es va iniciar amb la idea clara de convertir moltes dades disperses i heterogènies en informació útil, en definitiva – coneixement. Però aquesta conversió no s'havia de fer de qualsevol manera, calien uns resultats de qualitat que es poguessin consultar de manera ràpida i en qualsevol moment. Per fer-ho possible s'ha utilitzat una arquitectura que ha girat al voltant d'un Datawarehouse.

Amb el treball fet s'ha aconseguit una solució robusta. Per una banda és totalment escalable, necessita pocs recursos pel seu funcionament i al estar construïda sobre components oberts permet ser ampliada o adaptada als nous temps amb facilitat. Per altra, permet treballar no només amb les dades estàtiques pre-carregades, sinó que també amb les que són progressives en el temps.

En general el treball ha resultat molt positiu. S'han assolit tots els objectius marcats a l'inici, tant a nivell personal, com a nivell del projecte. El gran èxit aconseguit és l'enorme ampliació de coneixement personal en diferents àrees. S'ha pogut entendre de manera molt més clara els diferents aspectes d'un DW.

En la meva feina ja m'he trobat amb reptes d'optimització en el disseny de les BD o d'extracció de coneixement de les BD transaccionals clàssiques. Sobretot aquests últims m'han portat a concloure que per més optimitzada que sigui una base de dades, és pràcticament impossible tenir consultes no materialitzades que extreguin coneixement de les BD transaccionals de manera immediata. Els DW i el seu model multidimensional han estat tota una revelació. El seu joc de desnormalització-optimització en ROLAP obliga a conèixer molt a fons el funcionament dels DBMS, però també dóna uns resultats sorprenentment ràpids i eficients. A part, al tenir un enfocament d'extracció de coneixement des del principi, en comptes de només emmagatzemat d'informació, permet englobar més aspectes d'una organització que permetin visualitzar de de manera més fidel la realitat.

Un altre apartat que mereix ser mencionat és el d'ETL. S'ha vist la gran quantitat de feina que cal invertir per passar de dades heterogènies disperses a les estructurades que puguin ser tractades eficientment posteriorment. També s'ha après a utilitzar eines d'ETL que segurament seran molt útils en la vida professional.

La gran quantitat d'hores de feina dedicades a l'aprenentatge i implementació del model multidimensional i dels processos i eines ETL m'ha obligat a modificar el pla inicial i fer un pas enrere pel que fa a la implementació de l'eina OLAP d'exploració de dades. Sí que he pogut estudiar i entendre bé la teoria

del funcionament dels cubs OLAP, ja que està estrictament lligada amb el model multidimensional, però m'ha faltat temps per implementar de manera pràctica una eina d'exploració d'aquests cubs.

Pel que es refereix als informes, s'ha après a treballar amb una nova eina, Pentaho Report Designer, a part de la que ja coneixia (iReports de Jaspersoft). S'han vist les seves bondats i debilitats. Sobretot he quedat satisfet amb les possibilitats que ofereix juntament amb Pentaho BI Server. A partir d'aquí i més endavant consideraré aquesta solució en els projectes que desenvolupi.

Des d'una perspectiva de gestió del projecte també hi ha hagut diversos reptes a superar. La planificació i documentació, potser els més tediosos pels que tenim un perfil tècnic, han estat durs però també s'han vist els seus resultats positius. Per exemple, el seguiment correcte de la temporització i previsió de mesures de contingència han permès fer actuacions en moments crítics per la continuïtat del projecte.

Finalment, s'ha visualitzat la gran rigidesa en les estructures d'un DW. Aquesta manera més tradicional de tractament de dades comporta una bona dificultat en entorns més globalitzats, canviants i amb moltes dades arribant per tot arreu. Ara entenc més bé les noves tendències com el Big Data i estic impacient per aprendre'n més.

7. GLOSSARI DE TERMES

Base: en un disseny multidimensional són els diferents conjunts de Nivells que defineixin espais en què es poden col·locar les instàncies d'una Cel·la.

Business Intelligence (BI): o intel·ligència de negoci, consisteix en una sèrie de tècniques i eines que permeten donar sentit al gran volum de dades d'una organització amb la finalitat d'analitzar les operacions diàries i predir noves situacions.

cel·la: en un disseny multidimensional representa els mesuraments que fan referència al mateix esdeveniment, dins del mateix Fet.

Cel·la: en un disseny multidimensional representa el conjunt de cel·les del mateix Fet que estan associades a instàncies del mateix Nivell.

Database Management System (DBMS): és un aplicatiu que permet als usuaris i altres aplicatius interactuar amb les bases de dades, executant totes les funcions bàsiques.

Data Marts (DM): són repositoris de dades d'un DW que conglomeren informació especifica d'una àrea concreta de la organització.

Data Warehouse (DW): o magatzem de dades, és una base de dades que integra la informació provinent de diferents fonts heterogènies en una organització amb el propòsit de realitzar-hi consultes analítiques eficients i que sovint dóna suport a les diferents eines d'intel·ligència de negoci (BI).

Dimensió: en un disseny multidimensional representa un punt de vista que s'utilitza en l'anàlisi de dades.

Estrella: una tècnica de modelació del disseny multidimensional, amb un Fet central que es relaciona amb diverses Dimensions.

Extract, Transform, Load (ETL): o extreure, transformar i carregar, són un conjunt de processos que extreuen dades de les seves fonts originals, les netegen, adeqüen el format i les carregen en el DW.

Fet: en un disseny multidimensional representa un tema objecte d'anàlisi.

Floc de neu: en un disseny multidimensional, és una estrella totalment normalitzada. Es considera un error de disseny que rebaixa molt el rendiment de les consultes.

Grup Líder en Turisme Familiar (GLTF): empresa fictícia que encarrega el projecte.

Hardware (HW): o maquinari, es refereix a totes les parts físiques d'un sistema informàtic.

Mesura: en un disseny multidimensional representa un atribut d'una Cel·la.

Nivell: en un disseny multidimensional representa un conjunt d'instàncies d'una Dimensió que tenen la mateixa granularitat.

On-line Analytical Processing (OLAP): o processament analític en línia, són un conjunt de tècniques i eines que permeten als usuaris executar consultes analítiques multidimensionals (complexes) sobre un DW de manera ràpida i eficient.

Relational OLAP (ROLAP): eines de programari que reben consultes multi-dimensionals, les tradueixen al llenguatge SQL i les executen sobre un DBMS relacional.

Software (SW): o programari, es refereix als aplicatius que corren en un sistema informàtic.

Structured Query Language (SQL): llenguatge de programació dissenyat per gestionar les bases de dades relacionals. L'estàndard més utilitzat és l'especificació del 1999 (SQL'99).

Treball de Fi de Grau (TFG): és l'assignatura de caràcter pràctic dins de la qual s'inclou tot el projecte actual.

Unified Modeling Language (UML): llenguatge de modelat de sistemes de software, desenvolupat com una eina estàndard per visualitzar els seus dissenys.

8. BIBLIOGRAFIA

Bouman, Roland (2006). Pentaho Data Integration: Kettle turns data into business. [article en línia]. rpbouman.blogspot.com.es. [Data de consulta: 15 d'abril de 2015].
<http://rpbouman.blogspot.com.es/2006/06/pentaho-data-integration-kettle-turns.html>

Browning, Dave; Mundy, Joy (2001). Data Warehouse Design Considerations [article en línia]. Microsoft Corporation. [Data de consulta: 6 de març de 2015].
<https://technet.microsoft.com/en-us/library/aa902672(v=sql.80).aspx>

Di Doménico, Tomás (2010). Manual del Usuario de Spoon. [article en línia]. Pentaho.org. [Data de consulta: 14 d'abril de 2015].
<http://wiki.pentaho.com/display/EAIes/Manual+del+Usuario+de+Spoon>

Gavídia, Àngels; Serra, Montse; Abelló, Alberto; Samos, José; Vidal, Josep; Curto, Josep. (2012). Data warehouse: magatzems de dades i models multidimensionals (2a. ed.). Barcelona: Universitat Oberta de Catalunya.

Getz, Adam (2011). Benefits of a Data Warehouse [article en línia]. BI-Insider.com. [Data de consulta: 6 de març de 2015].
<http://bi-insider.com/portfolio/benefits-of-a-data-warehouse/>

Javlin Data Solutions (2015). ETL (Extract-Transform-Load) | Data Integration Info. [article en línia]. Javlin Data Solutions. [Data de consulta: 15 d'abril de 2015]. <http://www.dataintegration.info/etl>

Kazmi, Shehzad (2009). The Basics: OLTP and OLAP [article en línia]. About.com. [Data de consulta: 8 de març de 2015].
<https://analytiks.wordpress.com/2009/02/26/the-basics-oltp-and-olap/>

Mailvaganam, Hari (2007). Data Warehouse Project Management [article en línia]. DWreview.com. [Data de consulta: 7 de març de 2015].
<http://www.dwreview.com/Articles/Project_Management.html>

Mailvaganam, Hari (2007). Introduction to Metadata [article en línia]. DWreview.com. [Data de consulta: 7 de març de 2015].
<http://www.dwreview.com/Articles/Metadata.html>

Malinowski, Elzbieta (2008). Designing conventional, spatial, and temporal data warehouses : concepts and methodological framework : Data-centric systems and Applications. Estats Units: Springer.

Object Management Group, Inc. (2015). Introduction To OMG's Unified Modeling Language™ (UML®) [article en línia]. Object Management Group, Inc. [Data de consulta: 14 d'abril de 2015].
<http://www.omg.org/gettingstarted/what_is_uml.htm>

Oracle Corporation (2015). MySQL 5.1 Reference Manual. [article en línia]. Oracle Corporation. [Data de consulta: 26 de maig de 2015].
<http://dev.mysql.com/doc/refman/5.1/en/>

Pentaho community (2015). Latest Pentaho Data Integration (aka Kettle) Documentation [article en línia]. Pentaho.com. [Data de consulta: 26 de maig de 2015].
<http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>

Reh, John (2015). Pareto's Principle - The 80-20 Rule [article en línia]. About.com. [Data de consulta: 6 de març de 2015].
<http://management.about.com/cs/generalmanagement/a/Pareto081202.htm>

Rodríguez, José; Jové, Pere. (2010). Gestió de projectes (1a. ed.). Barcelona: Universitat Oberta de Catalunya.

Salazar, Bryan (2014). Variación Estacional o Cíclica – Ingeniería Industrial. [article en línia]. Ingenieriaindustrialonline.com. [Data de consulta: 14 d'abril de 2015].
<http://www.ingenieriaindustrialonline.com/herramientas-para-el-ingeniero-industrial/pron%C3%B3stico-de-ventas/variaci%C3%B3n-estacional-o-c%C3%ADclica/>

Steiner, Diethard (2013). Pentaho Report Designer: How to show the parameter display name in your report when it is different from the parameter value [article en línia]. Diethardsteiner.blogspot.com. [Data de consulta: 26 de maig de 2015].
<http://diethardsteiner.blogspot.com.es/2013/05/pentaho-report-designer-how-to-show.html>

Vogler, Raffael (2014). The Making of a Pretty Web Stats Report with Pentaho Report Designer 5 [article en línia]. Joyofdata.de. [Data de consulta: 26 de maig de 2015].
<http://www.joyofdata.de/blog/the-making-of-a-pretty-web-stats-report-with-pentaho-report-designer/>

9. ANNEXOS

9.1. ACCÉS A LES DADES

9.1.1. BASE DE DADES DEL DATAWAREHOUSE

La base de dades s'encapsula dins del nom **tfc_dw**, de la que depenen tots els altres components. Existeix un super-usuari amb tots els possibles privilegis, aquest no s'ha d'utilitzar per tasques diàries, més aviat per resoldre incidències extraordinàries i només es pot connectar des de l'ordinador local. Credencials:

Usuari	**root**
Contrasenya	**TFC2014**

S'ha configurat un canal de connexió pel super-usuari a través del MySQL Workbench.

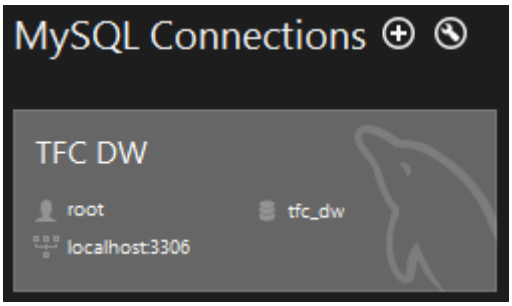


Figura 47. Connexió al DW amb super-usuari via Workbench

També s'han creat dos usuaris més per tasques més específiques.

Usuari **etl** s'utilitza en processos ETL. Té permisos de selecció, inserció i actualització de totes les taules de **tfc_dw**. La seva connexió només pot ser realitzada des d'ordinador local. Credencials:

Usuari	**etl**
Contrasenya	**TFC2014**

Usuari **normal** que serveix per generar informes i extreure coneixement general del DW. Només té permisos de selecció i la seva connexió pot ser local o remota indistintament.

Usuari	**normal**
Contrasenya	**TFC2014**

9.1.2. INFORMES

Tots els informes estan carregats en el servidor BI de Pentaho. L'accés es realitza via navegador web a través d'un dels dos usuaris definits.

El primer, **Admin**, és el super-usuari que té tots els permisos possibles. S'utilitza com a administrador d'altres usuaris i el qui carrega els informes en el servidor. Credencials:

Usuari	**Admin**
Contrasenya	**password**

El segon, **usuari**, només té la potestat per visualitzar tots els informes que li han sigut carregats prèviament. Credencials:

Usuari	**usuari**
Contrasenya	**TFC2014**

El circuit per accedir als informes és el següent:

1. Executar *start-pentaho* en l'escriptori.
2. Obrir el navegador en la URL *http://localhost:8080/pentaho/Login*.
3. Ingressar les credencials de l'usuari en qüestió.
4. Botó *Browse Files* ens porta a les carpetes dels usuaris. Cal seleccionar la nostra i obrir l'informe desitjat.

9.2. EXECUCIÓ DELS PROCESSOS ETL

Els fitxers amb processos ETL de Spoon es troben a la carpeta de l'escriptori de la màquina virtual anomenada *ETL*. Tots els processos han de ser executats segons un orde preestablert:

1. any.ktr
2. mes.ktr
3. estacions.ktr
4. establiment.ktr
5. municipi.ktr
6. detall_municipi.ktr
7. municipis_restants.ktr
8. marques_municipi.ktr
9. equipament.ktr
10. pernoctacions_2011.ktr
11. pernoctacions_2012.ktr
12. pernoctacions_2013.ktr
13. viatgers_2011.ktr
14. viatgers_2012.ktr
15. viatgers_2013.ktr
16. infraestructura_est_2011.ktr
17. infraestructura_est_2012.ktr
18. infraestructura_est_2013.ktr
19. infraestructura_pl_2011.ktr
20. infraestructura_pl_2012.ktr
21. infraestructura_pl_2013.ktr
22. infraestructura_policies.ktr
23. infraestructura_cens.ktr

9.3. INFORMES

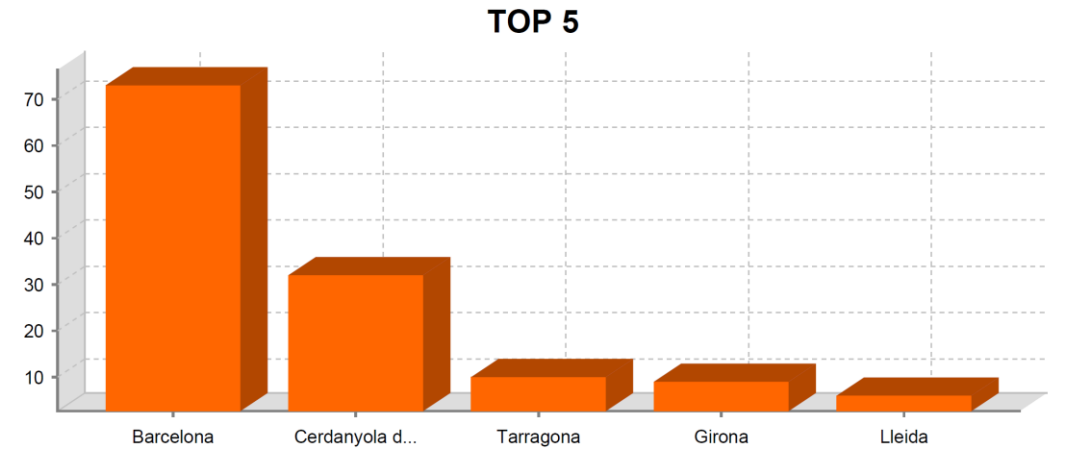
A continuació s'adjunta la primera pàgina de cada informe creat que permeten il·lustrar el resultat final. Els informes que necessitin una selecció de paràmetres prèvia per part de l'usuari s'han omplert amb paràmetres d'exemple aleatoris.

9.3.1. RÀNQUING DE MUNICIPIS PER CATEGORIA D'EQUIPAMENTS

Grup Líder en Turisme Familiar

Rànquing de municipis per categoria d'equipaments

Equipament: **Recerca**



Rànquing	Municipi	Número d'equipaments
1	Barcelona	73
2	Cerdanyola del Valls	32
3	Tarragona	10
4	Girona	9
5	Lleida	6
6	Castelldefels	5
7	Badalona	4
8	Roquetes	3
9	Terrassa	2
10	Hospitalet de Llobregat, L'	2
11	Viladecans	2
12	Alcarrs	1
13	Vic	1
14	Solsona	1
15	Sant Carles de la Rpita	1
16	Esplugues de Llobregat	1
17	Cabrils	1
18	Sant Esteve de la Sarga	1

1 / 2

Figura 48. Informe de rànquing de municipis d'equipaments de recerca

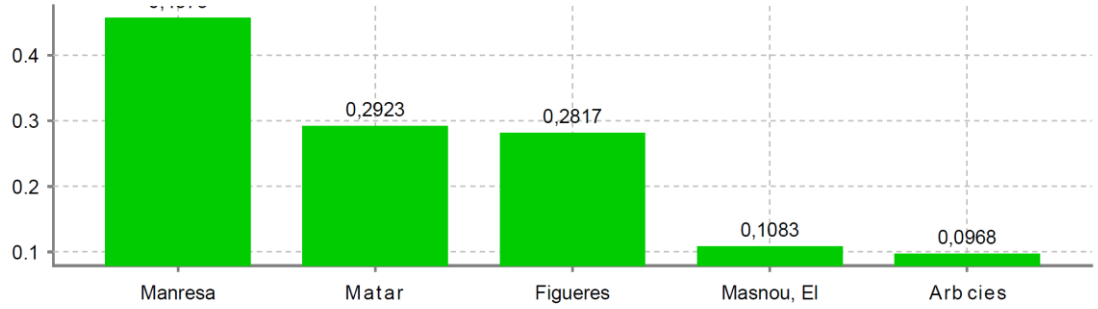
9.3.2. MÀXIM I MÍNIM D'EFECTIUS POLICIALS PER TIPOLOGIA D'ESTABLIMENT I MUNICIPI

Grup Líder en Turisme Familiar

Efectius policials per tipologia d'establiment i municipi

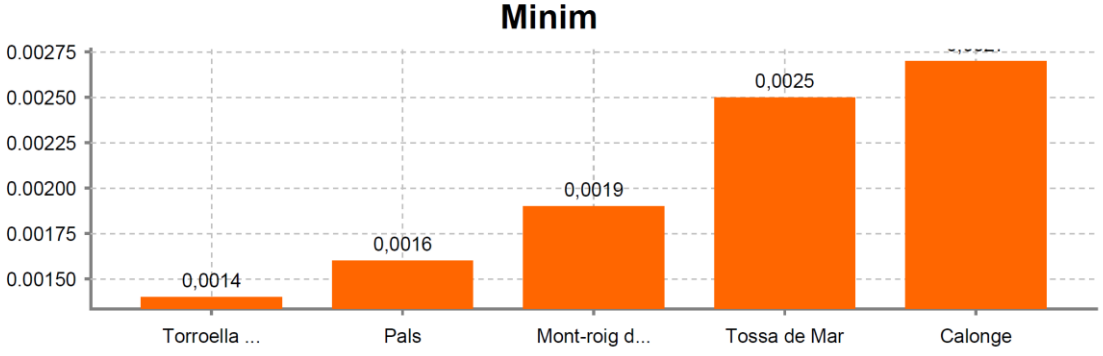
Any: **2011** Establiment: **camping**

Maxim



Municipi	Places	Policies	Rati
Manresa	234	107	0,4573
Matar	585	171	0,2923
Figueres	252	71	0,2817
Masnou, El	360	39	0,1083
Arbcies	93	9	0,0968

Minim



Municipi	Places	Policies	Rati
Torroella de Montgr	15.276	22	0,0014
Pals	8.196	13	0,0016
Mont-roig del Camp	14.010	27	0,0019
Tossa de Mar	7.863	20	0,0025
Calonge	8.580	23	0,0027

Figura 49. Informe de màxim i mínim d'efectius policials en càmpings en 2011

9.3.3. RATI DE POLICIES LOCALS PER HABITANT

Grup Líder en Turisme Familiar

Rati de policies locals per habitant

Any: **2012**

Municipi	Habitants	Policies	Habitants / Policia
Abella de la Conca	174	0	0
Abrera	11,870	21	565
ger	593	0	0
Agramunt	5,633	7	805
Aguilar de Segarra	250	0	0
Agullana	858	0	0
Aiguafreda	2,478	0	0
Aiguamrcia	906	0	0
Aiguaviva	774	0	0
Aitona	2,419	0	0
Alams, Els	751	0	0
Als i Cerc	376	0	0
Albags, L'	425	0	0
Albany	153	0	0
Albatrrec	2,113	0	0
Albesa	1,652	0	0
Albi, L'	860	0	0
Albinyana	2,367	0	0
Albiol, L'	459	0	0
Albons	714	0	0
Alcanar	10,658	16	666
Alcan	237	0	0
Alcarrs	8,755	12	730
Alcoletge	3,191	0	0
Alcover	5,143	0	0
Aldea, L'	4,530	0	0
Aldover	968	0	0
Aleixar, L'	926	0	0
Alella	9,610	20	481
Alfara de Carles	399	0	0
Alfarrs	3,077	0	0
Alfs	342	0	0

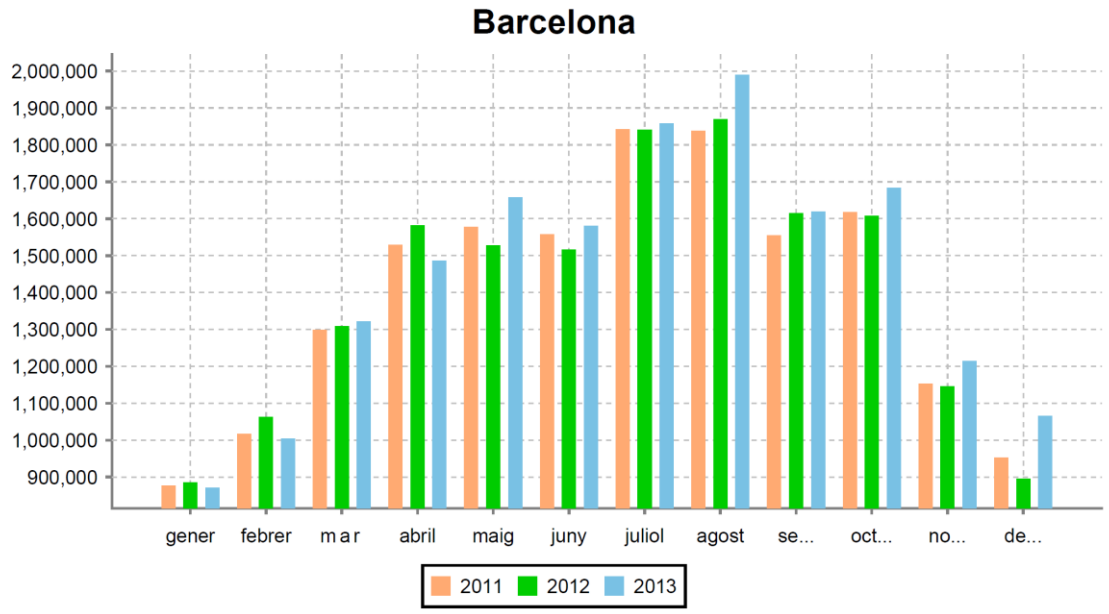
1 / 28

Figura 50. Informe de rati de policies locals per habitant en 2012

9.3.4. DISTRIBUCIÓ MENSUAL I ESTACIONALITAT DE LES PERNOCTACIONS

Grup Líder en Turisme Familiar

Distribució mensual i estacionalitat de les pernoctacions



Mes	2011	2012	2013
gener	877.352	885.470	871.107
febrer	1.016.489	1.063.321	1.003.787
mar	1.298.366	1.308.752	1.321.212
abril	1.528.372	1.581.859	1.485.955
maig	1.577.879	1.527.824	1.658.378
juny	1.556.905	1.516.095	1.580.935
juliol	1.842.939	1.841.379	1.858.624
agost	1.838.232	1.868.768	1.989.566
setembre	1.555.139	1.614.502	1.618.929
octubre	1.617.825	1.607.373	1.684.157
novembre	1.153.489	1.145.385	1.214.795
desembre	952.065	895.106	1.065.278

1 / 9

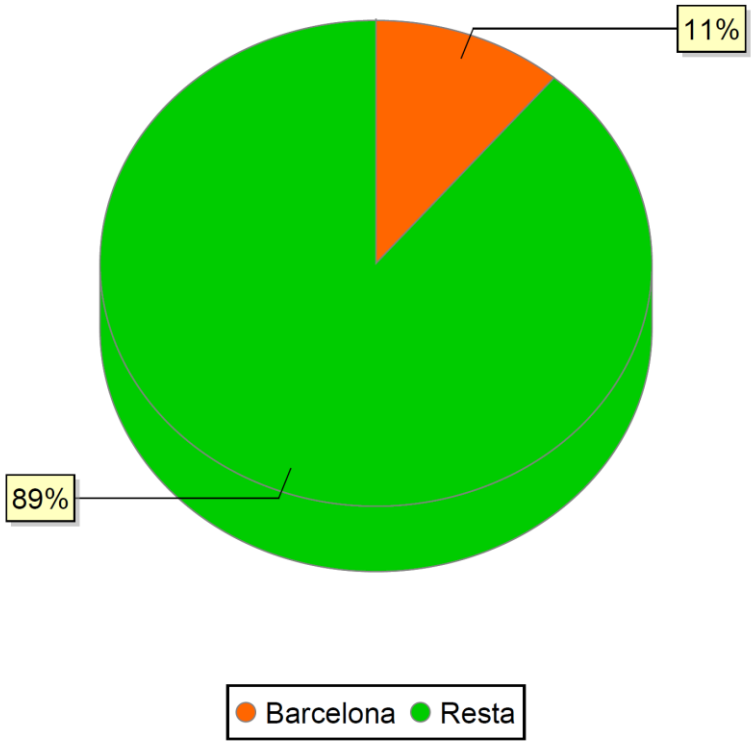
Figura 51. Informe de distribució mensual i estacionalitat de les pernoctacions

9.3.5. PERCENTATGE DE PERNOCTACIONS D'UN MUNICIPI SOBRE EL TOTAL

Grup Líder en Turisme Familiar

Percentatge de pernoctacions d'un municipi sobre el total

Any: **2013** Municipi: **Barcelona**



Pernoctacions del municipi:	**16.278.448**
Pernoctacions totals:	**145.845.646**
Percentatge:	**11,16%**

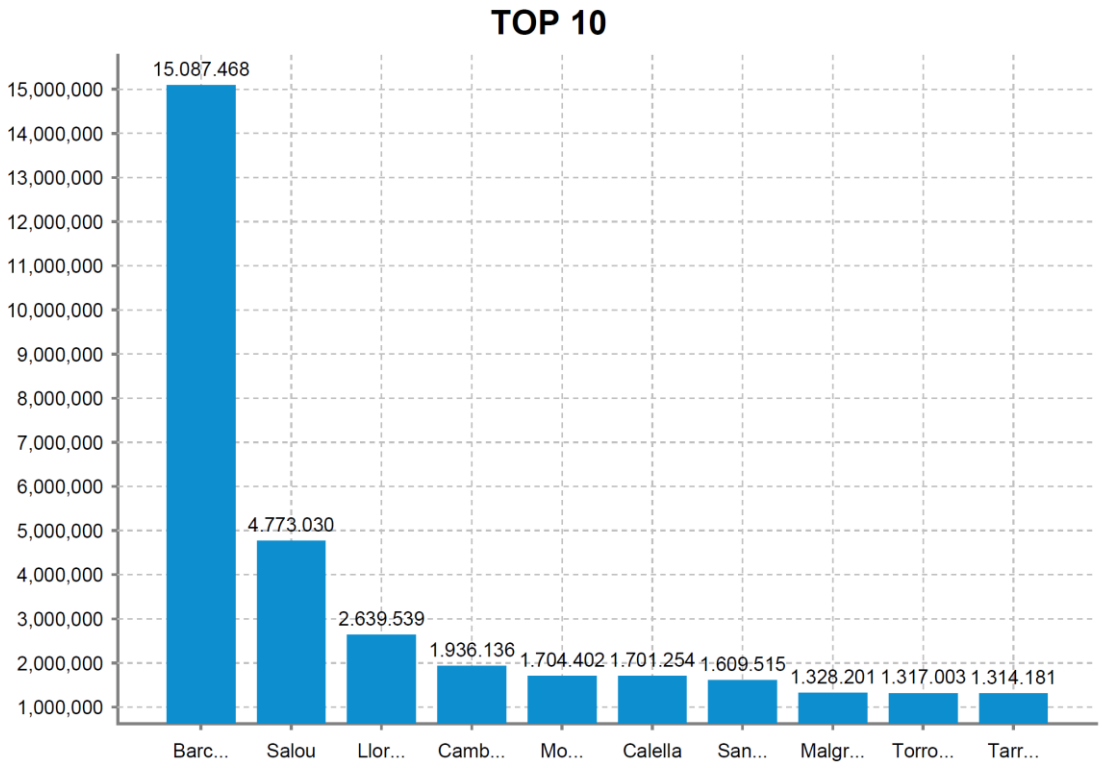
Figura 52. Informe de percentatge de pernoctacions de Barcelona en 2013 sobre el total

9.3.6. "TOP TEN" DE MUNICIPIS PER FRANJA DE PERNOCTACIONS

Grup Líder en Turisme Familiar

"Top ten" de municipis per franja de pernoctacions

Franja: **estiu**



Rànquing	Municipi	Pernoctacions
1	Barcelona	15.087.468
2	Salou	4.773.030
3	Lloret de Mar	2.639.539
4	Cambrils	1.936.136
5	Mont-roig del Camp	1.704.402
6	Calella	1.701.254
7	Santa Susanna	1.609.515
8	Malgrat de Mar	1.328.201
9	Torroella de Montgr	1.317.003
10	Tarragona	1.314.181

Figura 53. Informe de "top ten" de municipis per pernoctacions estiuenques

9.3.7. TOTAL DE PERNOCTACIONS ESTIMADES PER MUNICIPI

Grup Líder en Turisme Familiar

Total de pernoctacions estimades per municipi

Municipi	Pernoctacions per 2014
Abella de la Conca	98
Abrera	4.564
ger	32.980
Agramunt	3.864
Aguilar de Segarra	1.454
Agullana	654
Aiguafreda	0
Aiguamrcia	18.706
Aiguaviva	0
Aitona	0
Alams, Els	0
Als i Cerc	657
Albags, L'	418
Albany	19.834
Albatrrec	0
Albesa	0
Albi, L'	0
Albinyana	2.924
Albiol, L'	299
Albons	4.580
Alcanar	76.709
Alcan	0
Alcarrs	9.391
Alcoletge	0
Alcover	8.593
Aldea, L'	3.917
Aldover	337
Aleixar, L'	1.496
Alella	1.510
Alfara de Carles	861
Alfarrs	1.489
Alfs	6.982
Alforja	637

1 / 28

Figura 54. Informe de total de pernoctacions estimades per municipi per 2014

9.3.8. PROMIG DE VIATGERS PER TIPUS D'ESTABLIMENT I COMARCA

Grup Líder en Turisme Familiar

Promig de viatgers per establiment i comarca

Any: **2011** Establiment: **hotel**

Comarca	Promig de viatgers
Alt Camp	7.774,23
Alt Empord	216.334,77
Alt Peneds	18.100,32
Alt Urgell	17.271,91
Alta Ribagora	23.201,29
Anoia	57.359,99
Bages	111.541,75
Baix Camp	149.470,62
Baix Ebre	40.132,08
Baix Empord	207.044,19
Baix Llobregat	215.317,08
Baix Peneds	77.084,03
Barcelons	7.080.977
Bergued	14.532,07
Cerdanya	33.550,12
Conca de Barber	14.500,37
Garraf	121.601,06
Garrigues	7.984,55
Garrotxa	11.453,83
Girons	43.490,44
Maresme	755.246,33
Montsi	27.229,67
Noguera	21.988,31
Osona	165.228,1
Pallars Juss	8.322,38
Pallars Sobir	32.680,49
Pla d'Urgell	6.752,12
Pla de l'Estany	5.652,57
Priorat	4.990,15
Ribera d'Ebre	7.810,96
Ripolls	22.956,6
Segarra	9.179,85

1 / 2

Figura 55. Informe de promig de viatgers d'hotels en 2011 per comarques

9.3.9. PERCENTATGE D'OCUPACIÓ PER MARCA TURÍSTICA

Grup Líder en Turisme Familiar

Percentatge d'ocupació per marca turística

Any: **2013** Mes: **setembre**

Marca turística	Ocupació
Barcelona	73,23 %
Catalunya Central	47,38 %
Costa Barcelona	39,41 %
Costa Brava	23,73 %
Costa Daurada	38,57 %
Pirineus	5,63 %
Terres de l'Ebre	17,67 %
Terres de Lleida	14,07 %
Vall d'Aran	11,07 %

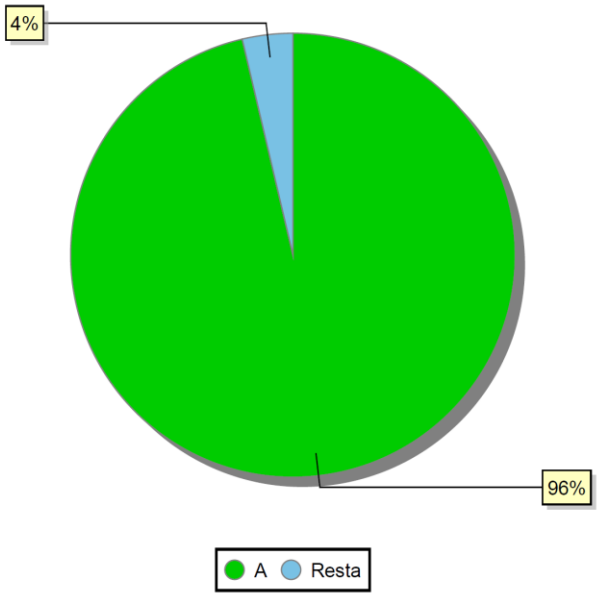
Figura 56. Informe de percentatge d'ocupació per marca turística en setembre del 2013

9.3.10. CATEGORITZACIÓ DE MUNICIPIS A/B/C

Grup Líder en Turisme Familiar

Categoritzacio de municipis A/B/C

Categoria: **A**



Municipi	Pernoctacions	Percentatge sobre total
Barcelona	15.087.468	23,7471 %
Salou	4.773.030	7,5126 %
Lloret de Mar	2.639.539	4,1545 %
Cambrils	1.936.136	3,0474 %
Mont-roig del Camp	1.704.402	2,6827 %
Calella	1.701.254	2,6777 %
Santa Susanna	1.609.515	2,5333 %
Malgrat de Mar	1.328.201	2,0905 %
Torroella de Montgr	1.317.003	2,0729 %
Tarragona	1.314.181	2,0685 %
Castell-Platja d'Aro	1.288.884	2,0287 %
Tossa de Mar	1.199.364	1,8878 %
Blanes	1.145.526	1,803 %
Sant Pere Pescador	1.100.301	1,7318 %
Pineda de Mar	901.325	1,4187 %

1 / 7

Figura 57. Informe de municipis de la categoria A