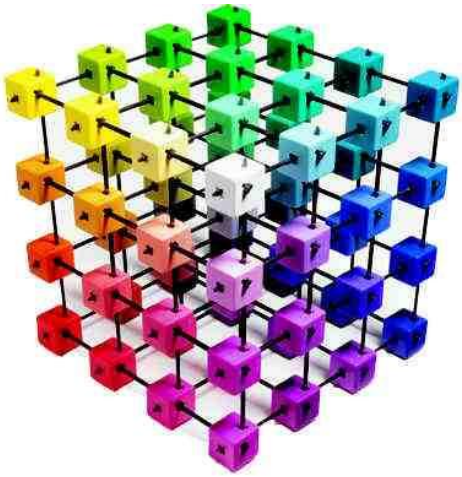
TREBALL FINAL DE CARRERA

MEMÒRIA



Construcció i explotació d'un magatzem de dades per a l'anàlisi del sistema de prestacions socials

Autor: Iván González Díez
ETIS
Consultor: Pere Juanola Juanola

Universitat Oberta de Catalunya

Data de lliurament: 13 de juny de 2012

Dedicatòria

Aquest treball el vull dedicar a aquelles persones que mai han deixat de donar-me ànims i recolzament per finalitzar els estudis, tot i que hi han hagut èpoques dures en les que mantenir la concentració per estudiar ha estat una tasca complicada. Família, amics, companys de feina i, en especial, al meu avi. Tot i que ja no es troba entre nosaltres, segur que estaria molt orgullós de saber que he aconseguit acabar aquells estudis que vaig començar fa molt de temps, quan encara estàvem junts.

Resum

El present projecte consisteix en la construcció i explotació d'un magatzem de dades per l'OADT (Organisme d'Anàlisi del Departament de Treball). Aquest magatzem ha de permetre determinar la sostenibilitat i conèixer l'evolució del sistema de prestacions socials.
El punt de partida són diversos fitxers amb informació tributària i de població, de les diferents CCAA (Comunitats Autònomes) d'Espanya. A partir d'aquestes dades, es realitza l'anàlisi previ, el disseny conceptual, lògic i físic per acabar construint el cub multidimensional (OLAP) que és la base de la implementació del projecte .
Mitjançant diferents algorismes en llenguatge de programació PL/SQL, es realitza la transformació i càrrega de la informació font a la base de dades. A partir d'aquí, es crea l'àrea de negoci i s'implementen les diferents consultes i informes que serveixen per donar resposta als requeriments del client final.

Paraules clau

Magatzem de dades, Data warehouse, OLAP, Organisme d'Anàlisi del Departament de Treball, SQL, PL/SQL, ETL, Oracle, Discoverer.

Índex de Continguts

Índex de Taules i Figures

1. Introducció

Amb la present documentació es descriu la memòria final d'un Treball Final de Carrera dels estudis d'Enginyeria Tècnica d'Informàtica de Sistemes, dintre de l'àrea de Magatzem de Dades.

1.1. Justificació del TFC i context en el que es desenvolupa: punt de partida i aportació del TFC

Aquest projecte es desenvolupa per donar resposta als requeriments de L'OADT (Organisme d'Anàlisi del Departament de Treball) que, davant la situació de crisi actual i el constant increment del nombre de persones en atur, ha decidit fer un estudi per determinar la sostenibilitat del sistema de prestacions socials i conèixer l'evolució d'aquestes prestacions. Per portar a terme el projecte és necessària la creació d'un magatzem de dades.

Amb el producte final s'aconsegueix tenir una sèrie de processos automatitzats que realitzen les tasques de transformació i càrrega de les dades aportades pel client. Aquestes dades, un cop organitzades al magatzem de dades, es podran explotar amb l'eina Oracle Discoverer, per aportar el coneixement requerit pel client i es deixa la possibilitat de creació de noves consultes i informes amb relativa facilitat.

1.2. Objectius del TFC

L'objectiu principal del projecte és adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible en una base de dades transaccional.
A partir d'aquest objectiu general es poden identificar una sèrie d'objectius més concrets:

- Aprofundir en el coneixement sobre els magatzem de dades per donar resposta als requeriments del client.
- Familiaritzar-se en el treball amb bases de dades Oracle, emprant SQLDeveloper.
- Realitzar els processos de modificació i càrrega de dades emprant PL/SQL.
- Obtenir experiència en el desenvolupament d'àrees de negoci, emprant Oracle Discoverer Administrator.
- Obtenir experiència en la implementació de consultes i informes per l'explotació del magatzem de dades, emprant Oracle Discoverer Desktop.

A part dels temes tècnics relacionats amb la creació del producte, s'obté experiència en el desenvolupament complet de projectes, partint de les dades aportades pel client i de les seves necessitats. És necessari analitzar el que es demana i les dades de què es disposa, realitzar una planificació de les tasques a realitzar, dissenyar i desenvolupar el model de dades necessari i els processos que faran que les dades estiguin carregades al model, realitzar una memòria i una presentació virtual del projecte.

1.3. Enfocament i mètode seguit

L'enfocament i el mètode seguit per desenvolupar aquest TFC es pot esquematitzar amb les fases més comuns per les que passa un projecte. S'expliquen a continuació:

1. **Anàlisi preliminar de requeriments:** En aquesta fase s'estudien els requeriments del client i es realitza un primer document de requisits. A més, s'inclou una aproximació dels elements d'anàlisi que conformaran el cub multidimensional (OLAP), el nombre d'informes aproximat que s'implementaran, amb el contingut dels mateixos, i la planificació estimada de les diferents tasques a realitzar per a dur a terme el projecte.

2. **Instal·lació del suport físic:** En aquesta fase es realitza la instal·lació, tant a nivell de hardware com de software, dels elements necessaris per la implementació del magatzem de dades i les eines d'explotació. Es fa servir la màquina virtual proporcionada amb el programari ja instal·lat.

3. **Disseny del model de dades:** A partir del document de l'anàlisi preliminar es realitza un anàlisi detallat de requeriments. Es realitza el disseny a nivell conceptual, lògic i físic segons l'anàlisi realitzat.

4. **ETL:** Aquesta fase d'ETL (Extraction, Transformation, Loading) s'encarrega d'extreure les dades de les diferents fonts, manipular-les perquè estiguin en el format adequat i carregar-les a les taules definitives de la base de dades.

5. **Disseny del cub OLAP:** A partir de tota la informació que s'obté de l'anàlisi, es dissenya el cub OLAP que dóna resposta a les necessitats imposades pel client, podent realitzar les consultes i informes indicats.

6. **Desenvolupament de les eines de consulta per l'usuari final:** En aquesta fase es defineix quina aplicació s'utilitzarà per la realització de les consultes i informes. Per aquest projecte s'utilitza Oracle Discoverer.

7. **Control del rendiment:** Es realitzen diverses proves de rendiment, sobretot a l'hora d'executar els processos ETL i de realitzar les consultes i informes. És important no realitzar més operacions de transformació i càrrega de les necessàries, utilitzar els índexs adients a les taules de la base de dades i incloure als informes només aquells elements que siguin necessaris.

8. **Comprovació de la qualitat del producte final:** En aquesta fase el client prova el producte per determinar si compleix les seves necessitats.

9. **Posada en marxa del producte en producció:** Es posa en funcionament el producte en producció.

10. **Manteniment:** És necessari realitzar un manteniment del producte.

11. **Millores:** És habitual que els requeriments dels clients canviïn conforme passa el temps i que s'hagin de realitzar ampliacions del producte.

Les quatre últimes fase no és possible portar-les a terme, ja que el producte final no es posa en producció.

1.4. Recursos utilitzats

1.4.1. Hardware

A casa es disposa de dos ordinadors sobretaula i un portàtil:

- AMD Athlon 64bits 3000+ amb 3GB de RAM, Windows 7 Enterprise.
- HP DC6005 AMD PhenomIIX2 B55 3Ghz amb 3GB de RAM, Windows XP Professional Service Pack 3.
- Toshiba Portege M400-167 Core 2 Duo T7400 amb 2GB de RAM, Windows XP Professional Service Pack 3.

A la feina es disposa d'un ordinador sobretaula:

- HP Compaq 6000 Pro SFF Core 2 Duo E8500 amb 2GB de RAM, Windows XP Professional Service Pack 3.

Majoritàriament es fan servir els equips de casa, però algunes tardes o els caps de setmana es fan servir el de la feina o portàtil.

1.4.2. Software

El software que es fa servir per la implementació del projectes és:

- Microsoft Project 2010 per la planificació del TFC.
- Microsoft Word 2003/2007 per la realització de la memòria del projecte.
- Microsoft Powerpoint 2003/2007 per la realització de la presentació.
- Oracle 10XE i SQL Developer per treballar amb la base de dades.
- Magic Draw UML Personal Edition i ArgoUML per treballar amb els diagrames E/R i la seva transformació en disseny lògic.

1.5. Planificació del projecte

Per realitzar la temporització i una planificació de manera adequada, és fonamental tenir en compte el calendari publicat al pla d'estudis de l'assignatura. És el següent:

Títol	Inici	Entrega
PAC 1	01/03/2012	16/03/2012
PAC 2	17/03/2012	20/04/2012
PAC 3	21/04/2012	25/05/2012
Lliurament Final	26/05/2012	13/06/2012
Debat Virtual	25/06/2012	28/06/2012

Taula 1.- Calendari del pla d'estudis

La idea inicial és treballar en el TFC durant 3 hores diàries els dies laborables. Tot i així, aquest càlcul s'altera en funció de la càrrega de treball a la feina o altres factors externs. Per això, alguns caps de setmana també es treballa en el projecte.

1.5.1. Descripció de tasques

PAC 1: Pla de treball i anàlisi preliminar de requeriments

- Descripció del projecte a partir de la documentació donada pel consultor.
- Definició dels objectius i tasques que comporten el projecte.
- Realització del pla de treball amb la planificació estimada de les diferents tasques a realitzar per dur a terme el projecte.
- Realització de l'anàlisi preliminar (no detallat) amb l'enumeració i breu descripció dels elements d'anàlisi identificats (dimensions, atributs, indicadors, etc.) que estaran disponibles per als usuaris, i el nombre d'informes aproximat que s'implementaran i contingut dels mateixos.
- Analitzar les fonts de dades operacionals proporcionades que serviran per carregar cadascun dels elements d'anàlisi.
- Revisió del document abans de l'entrega final.

PAC 2: Anàlisi de requeriments i disseny conceptual i tècnic

- Anàlisi detallat de requeriments basat en l'anàlisi preliminar realitzat.
- Disseny amb la descripció del model dimensional que donarà suport a les necessitats dels usuaris, segons l'anàlisi realitzat.
- Disseny dels procediments d'extracció de dades a alt nivell (processos, pseudocodi, etc.).
- Preparació del programari (Oracle) per la posterior utilització.

PAC 3: Implementació

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

Entrega final

- Redacció de la memòria final del projecte.
- Creació de la presentació virtual (Power Point) amb àudio.

- Revisió (i modificació, si cal) d'aquells punts que es creguin convenients abans de l'entrega final.

1.5.2. Fites

A part de les dates d'entrega de les activitats obligatòries (PACs) és recomanable afegir que, en la mesura del possible, es fa l'entrega d'esborranys de les PACs. Aquesta entrega es fa aproximadament una setmana abans de l'entrega definitiva perquè el consultor pugui orientar si el desenvolupament de la mateixa és adequat.

1.5.3. Calendari previst

A continuació s'observa una taula amb detall de les tasques anteriorment descrites, amb la seva durada, data d'inici i data fi, durada dels dies dels principals esdeveniments i la precedència que necessita cadascuna de les tasques.

ID	Nom de la tasca	Durada	Inici	Fi	Predecessores
1	**1. Inici del projecte**	**5**	**29/02/2012**	**04/03/2012**	
2	1.1 Descarregar, estudi material i recerca bibliogràfica	5	29/02/2012	04/03/2012	
3	1.2 Lectura enunciat TFC	1	04/03/2012	04/03/2012	2
4	**2. Elaboració del esborrany de la PAC 1**	**10**	**05/03/2012**	**14/03/2012**	
5	2.1 Instal·lació Microsoft Project	1	05/03/2012	05/03/2012	
6	2.2 Descarregar i llegir PACs anys anteriors	2	05/03/2012	06/03/2012	
7	2.3 Elaboració esborrany del Pla de treball	4	05/03/2012	08/03/2012	
8	2.5 Elaboració esborrany de l'Anàlisi preliminar de requeriments	6	08/03/2012	13/03/2012	
9	2.6 Lliurament de l'esborrany de la PAC 1	1	14/03/2012	14/03/2012	8
10	**3. Elaboració PAC 1**	**2**	**15/03/2012**	**16/03/2012**	
11	3.1 Revisió i modificacions sobre l'esborrany de la PAC 1	1	15/03/2012	15/03/2012	
12	3.2 Elaboració Pla de treball definitiu	2	15/03/2012	16/03/2012	
13	3.3 Elaboració de l'Anàlisi preliminar de requeriments definitiu	2	15/03/2012	16/03/2012	
14	3.4 Lliurament de la PAC 1	1	16/03/2012	16/03/2012	12;13
15	**4. Elaboració del esborrany de la PAC 2**	**28**	**17/03/2012**	**13/04/2012**	
16	4.1 Nova lectura enunciat TFC	1	17/03/2012	17/03/2012	
17	4.2 Redacció de la introducció del TFC	2	18/03/2012	19/03/2012	
18	4.3 Realització d'esborrany de l'Anàlisi de requeriments	5	20/03/2012	24/03/2012	

19	4.4 Realització del model de dades	18	25/03/2012	11/04/2012	
20	4.5 Realització esborrany del disseny tècnic (conceptual i físic)	18	25/03/2012	11/04/2012	
21	4.6 Revisió de documentació	2	12/04/2012	13/04/2012	
22	4.7 Lliurament esborrany PAC 2	1	13/04/2012	13/04/2012	17;18;19;20;21
23	**5. Elaboració de la PAC 2**	**7**	**14/04/2012**	**20/04/2012**	
24	5.1 Revisió de les correccions sobre l'esborrany	1	14/04/2012	14/04/2012	
25	5.2 Elaboració del document definitiu de l'anàlisi de requeriments i del disseny tècnic	6	15/04/2012	20/04/2012	
26	5.3 Lliurament de la PAC 2	1	20/04/2012	20/04/2012	25
27	**6. Elaboració de l'esborrany de la PAC 3**	**29**	**21/04/2012**	**19/05/2012**	
28	6.1 Instal·lació d'Oracle 10XE y SQL developer	1	21/04/2012	21/04/2012	
29	6.2 Familiarització amb el software	1	22/04/2012	22/04/2012	
30	6.3 Nova lectura detallada de l'enunciat	1	23/04/2012	23/04/2012	
31	6.4 Construcció del magatzem de dades (bases de dades, càrregues, etc)	24	23/04/2012	16/05/2012	
32	6.5 Construcció dels informes i anàlisi de la informació	24	23/04/2012	16/05/2012	
33	6.6 Elaboració de l'esborrany de la PAC 3	8	12/05/2012	19/05/2012	
34	6.7 Redacció del glossari del TFC	8	12/05/2012	19/05/2012	
35	6.8 Lliurament esborrany de la PAC 3	1	19/05/2012	19/05/2012	33
36	**7. Elaboració de la PAC 3**	**6**	**20/05/2012**	**25/05/2012**	
37	7.1 Revisió de les correccions de l'esborrany	1	20/05/2012	20/05/2012	
38	7.2 Depuració dels informes	5	21/05/2012	25/05/2012	
39	7.3 Elaboració del document definitiu de la PAC 3	5	21/05/2012	25/05/2012	
40	7.4 Elaboració de la presentació	6	20/05/2012	25/05/2012	
41	7.5 Lliurament de la PAC 3	1	25/05/2012	25/05/2012	39
42	**8. Elaboració de la memòria definitiva**	**19**	**26/05/2012**	**13/06/2012**	
43	8.1 Creació del document final	11	26/05/2012	05/06/2012	
44	8.2 Revisió i/o modificacions de la presentació	8	06/06/2012	13/06/2012	
45	8.3 Entrega final del projecte	1	13/06/2012	13/06/2012	43;44
46	**9. Debat**	**4**	**25/06/2012**	**28/06/2012**	
47	9.1 Temps per debatre	4	25/06/2012	28/06/2012	

Taula 2.- Relació de tasques

Nota: a la taula s'indica la durada incloent els caps de setmana, ja que molts d'ells es treballa en el TFC, però a la planificació s'indica que idealment es realitza de dilluns a divendres.

1.5.4. Diagrama de Gantt

Del calendari anterior s'obté els següent diagrama de Gantt, on es pot observar gràficament com evoluciona el projecte en relació al temps establert.

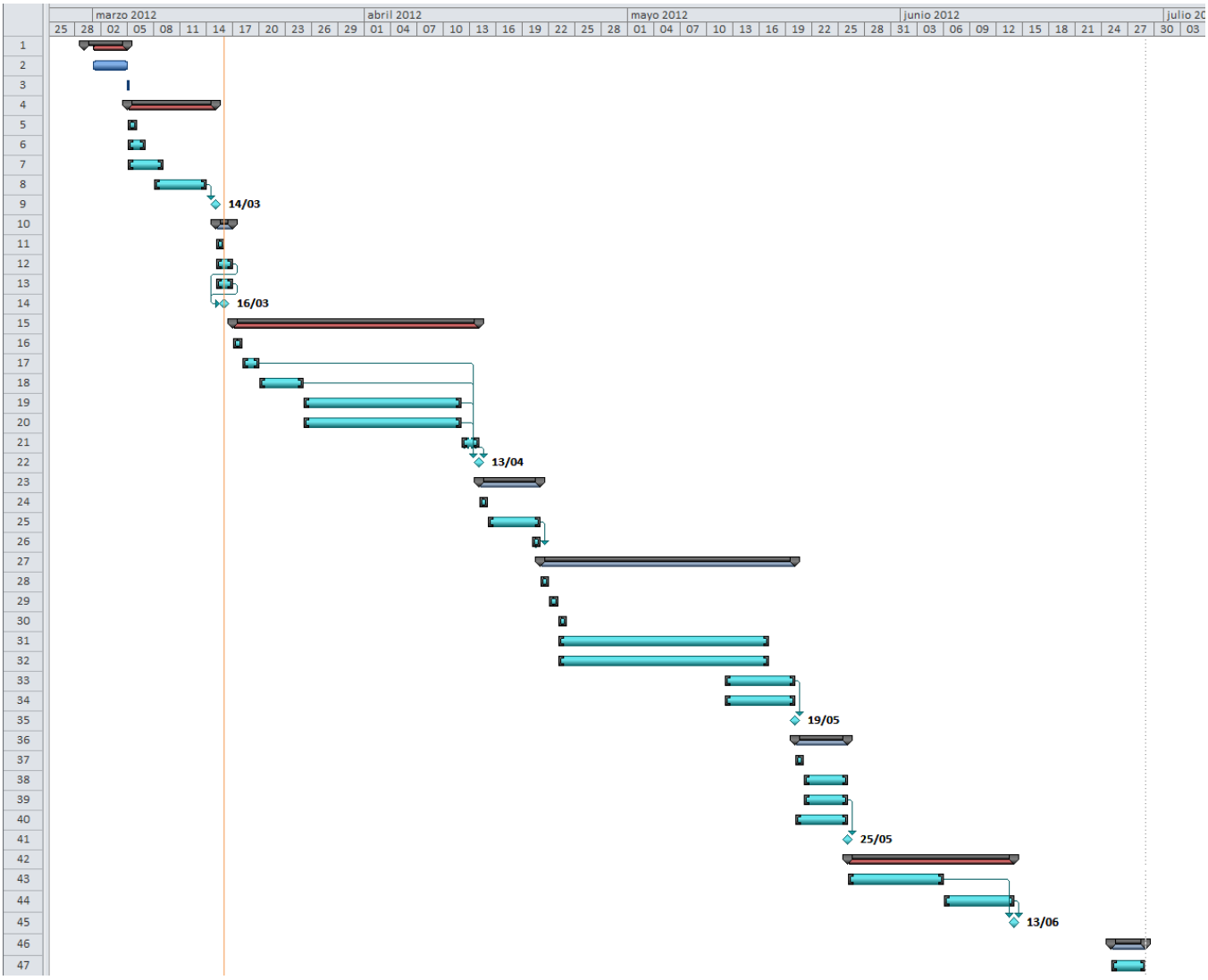


Figura 1.- Diagrama de Gantt

1.5.5. Incidències i riscos – Pla de contingències

Durant la realització de qualsevol projecte poden sorgir diversos problemes a nivell laboral o personal que poden fer que la realització de les tasques no es puguin seguir com estava determinat. És per això que, abans de començar el projecte, cal tenir un pla alternatiu en cas que succeeixin.

Avaria al punt de treball

El TFC es realitza amb els equips de sobretaula de casa principalment. Es disposa d'un portàtil per utilitzar els cap de setmana que sigui necessari. En cas que un dels dos equips falli, hi ha la possibilitat de treballar en l'altre, ja que es fan còpies de seguretat en un disc dur extern. A més a més, es pot fer servir l'equip de la feina, que en un moment de necessitat es pot utilitzar per ús particular.

Compromisos

En principi no es preveu cap compromís extraordinari, potser un canvi de residencia habitual, però sense definir.

Motius de salut

En cas de malaltia no greu, es pot continuar treballant des de casa. En cas de malaltia greu es poden aprofitar els caps de setmana o ampliar la jornada d'estudi de dilluns a divendres, un cop recuperat.

1.5.6. Desviació respecte la planificació prevista

Anteriorment s'ha explicat la planificació prevista, els possibles riscos i el corresponent pla de contingències. Finalitzat el projecte, es pot afirmar que la planificació ha estat adequada i s'han portat a terme pràcticament la totalitat de les fites en els terminis establerts. Les úniques fites que no s'entreguen a la data prevista són l'esborrany de la PAC 1, que s'entrega un dies més tard, i l'esborrany de la PAC 3 que no és possible entregar-ho.

Aquestes variacions d'entrega no han afectat a les dates d'entrega dels documents definitius de cada PAC ni a l'entrega final.

1.6. Productes obtinguts

Amb la realització del projecte es generen diferents documents i fitxers. A continuació, s'indiquen quins productes són, segons les fases que determina cada PAC:

Pla de treball

Document on s'indica la planificació per la realització del projecte, segons els objectius establerts, amb el corresponent pla de contingències i recursos dels que es disposa. Forma part de la PAC 1.

Anàlisi preliminar de requeriments

Document on s'analitzen les fonts de dades proporcionades pel client, s'identifiquen els elements d'anàlisi i on es descriuen els informes a realitzar. Forma part de la PAC 1.

Anàlisi de requeriments

Document on es detallen, de manera definitiva, els elements de l'anàlisi preliminar i que constitueixen la fase d'anàlisi. Forma part de la PAC 2.

Disseny

Document on es detalla la fase de disseny conceptual, lògic i físic del model dimensional que donarà suport a les necessitats dels usuaris, segons l'anàlisi realitzat i el disseny dels procediments ETL. Es realitza una primera aproximació de la metodologia d'implementació de les consultes. Forma part de la PAC 2.

Implementació

Creació de la base de dades, càrrega de les dades amb les corresponents transformacions, construcció dels informes i anàlisi de la informació. Forma part de la PAC 3.

Memòria final

És el present document. Inclou una síntesi de tot el procés de realització del projecte, incloent la major part de la informació continguda a cada una de les entregues parcials. Forma part del lliurament final.

Presentació virtual

Document que serveix per exposar de manera visual, esquemàtica i sense entrar en detall, el contingut de la memòria.

Producte

Conté tots els fitxers que es creen per la implementació del projecte. Són els següents:

- *creataules.sql*: script SQL que crea totes les taules necessàries, tant les temporals com les definitives, i les vistes materialitzades que s'utilitzen al magatzem de dades.

- *carregaCTLs.bat*: fitxer d'execució per lots que realitza la carrega de les dades tributàries de totes les CCAA (Comunitats Autònomes) a la taula temporal. Consta de línies d'execució de sqlloader per cada fitxer font.

- *carregaPobCA.bat*: fitxer d'execució per lots que realitza la carrega de les dades de població i població activa de totes les CCAA a les taules temporals. Consta de línies d'execució de sqlloader per cada fitxer font.

- *carregaPrestacio.sql*: script sql que efectua l'acondicionament de les dades tributàries i omple la taula definitiva PRESTACIO.

- *carregaCursorsPobCA.sql*: script sql que realitza l'acondicionament de les dades de població i de percentatge de població activa i omple la taula definitiva POBLACIO.

- *modificaValorsPobCA.sql*: script sql que realitza la modificació del número d'habitants, realitzant les mitjanes entre la població a 1 de gener de l'any i l'1 de gener de l'any següent.

- *carregaProjeccio.sql*: script sql que realitza càlculs matemàtics i càrrega de resultats a una taula temporal, per realitzar la projecció dels indicadors requerida pel client. S'explica amb detall a l'apartat de les consultes.

- *ivangd_mv.vdi*: disc dur de la màquina virtual on es troba la base de dades i tots els fitxers utilitzats en la implementació.

- *llibreOADT.dis*: llibre de treball.

A part dels fitxers anteriors, s'inclouen tots els fitxers proporcionats per l'OADT i que són la font d'informació per posteriorment realitzar les consultes i informes. Hi ha fitxers amb informació tributària i fitxers amb informació de població i població activa de cada CA.

1.7. Altres capítols de la memòria

1.7.1. Anàlisi

En aquest capítol s'estudien les dades origen i s'analitzen les transformacions necessàries per poder ser útils al magatzem de dades. S'estudia la viabilitat del projecte en quant al volum de dades que es tracten. S'avaluen els objectius que s'han d'aconseguir, tenint en compte les consultes i informes requerits pel client. Es fa el disseny conceptual del model de dades multidimensional que serà necessari per la implementació del projecte. S'avaluen els terminis del desenvolupament del projecte.

1.7.2. Disseny

En aquest capítol es realitza el disseny lògic, on es defineixen les taules que s'utilitzaran al magatzem de dades amb les corresponents claus primàries, claus foranes i relacions entre les taules. Es realitza el disseny físic a partir del disseny lògic, utilitzant índexs per accelerar les consultes posteriors. S'indica com seran els processos ETL. Es dissenyen i descriuen els informes creats. S'indiquen els requeriments, a nivell de hardware i software, pel desenvolupament i execució del producte final.

2. Anàlisi

2.1. Estudi inicial

Davant la situació de crisi actual i el constant increment del nombre de persones en atur, l'OADT ha decidit fer un estudi per determinar la sostenibilitat del sistema de prestacions socials i conèixer l'evolució d'aquestes prestacions. L'OADT encarrega la creació d'un magatzem de dades, mitjançant el qual es pugui obtenir certa informació explotant les dades mitjançant eines de Business Intelligence.

El client proporciona informació sobre prestacions (Salaris, Pensions, Atur, etc), les retencions realitzades i el nombre de persones que rep cada prestació de totes les CCAA, exceptuant el País Basc i Navarra que tenen un règim foral propi. A més, facilita dades de població i persones en actiu per CCAA i any.

L'OADT proporciona tota la informació en els següents fitxers en format CSV (de l'anglès, Comma-Separated Values):

- *Población por CCAA xxxx*: Un arxiu per cada any (del 2005 al 2011) que guarda la quantitat de població de cada CA. Tots mantenen la mateixa estructura i tenen la coma com a delimitador de camp. Els camps que contenen són:
    - Nom de la CA. Guarda els noms de les comunitats i, a més, hi ha un registre que guarda els totals de sumar els habitants de totes les CCAA. El camp és alfanumèric, amb la longitud màxima de 28 caràcters.
    - Nombre d'habitants de sexe masculí. Camp alfanumèric, amb una longitud màxima de 13 caràcters. En realitat és un valor numèric però entre cometes.
    - Nombre d'habitants de sexe femení. Camp alfanumèric, amb una longitud màxima de 13 caràcters. En realitat és un valor numèric però entre cometes.

- *Porc Población Activa*: Arxiu que guarda els percentatges de població activa que hi ha a les diferents CCAA (entre els anys 2005 i 2010). La coma és el delimitador de camp.
  Els camps que conté són:
    - Nom de la CA. Guarda els noms de les comunitats i, a més, hi ha un registre que guarda els totals de sumar els percentatges de totes les CCAA. El camp és alfanumèric, amb la longitud màxima de 28 caràcters.
    - Un camp per cada any del 2005 al 2010. Guarda el percentatge de població activa de cada CA i cada any. Camp alfanumèric, amb una longitud màxima de 6 caràcters. En realitat és un valor numèric però entre cometes.

- *Tributación XXXXXXX*: Un arxiu de cada CA, amb informació del nombre de persones, el total de retribucions i les retencions per any i tipus de perceptor (assalariat, pensionista, aturat, assalariat/pensionista, assalariat/aturat, pensionista/aturat, assalariat/pensionista/aturat). Els camps que contenen són:
    - Camp amb la CA a la que pertanyen les dades, els noms de les prestacions analitzades i els totals. El camp és alfanumèric, amb la longitud màxima de 25 caràcters.

- Un camp per cada tipus de perceptor i cada any analitzat. Camp alfanumèric, amb una longitud màxima de 18 caràcters. En realitat és un valor numèric però entre cometes.

El client vol la informació dintre d'una temporalitat a nivell d'any i poder consultar-la de forma agregada per CA, tipus de perceptor i tipus de retribució. S'indica també que, per que les dades siguin més realistes, el nombre d'habitants s'ha de calcular com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent.

Amb tota aquesta informació i amb la necessitat de fer les explotacions de dades que requereix el client, es construeix un magatzem de dades per guardar la informació i treballar després amb les eines de Business Intelligence.

2.2. Tractament de la font de dades

Les dades incloses als fitxers CSV es traspassen a taules de la base de dades Oracle. Es fan servir processos ETL, que s'utilitzen per moure dades des de múltiples fonts, reformatar-los i netejar-los, per posteriorment carregar-los a una base dades, data mart o data warehouse per analitzar, o a altre sistema operacional per ajudar a un procés de negoci.
Es tenen en compte les diferències que hi ha entre els fitxers aportats per les CCAA, ja que no segueixen un Standard degut a que s'ha extret la informació de diferents sistemes.

A la següent imatge es pot veure com és el procés ETL:

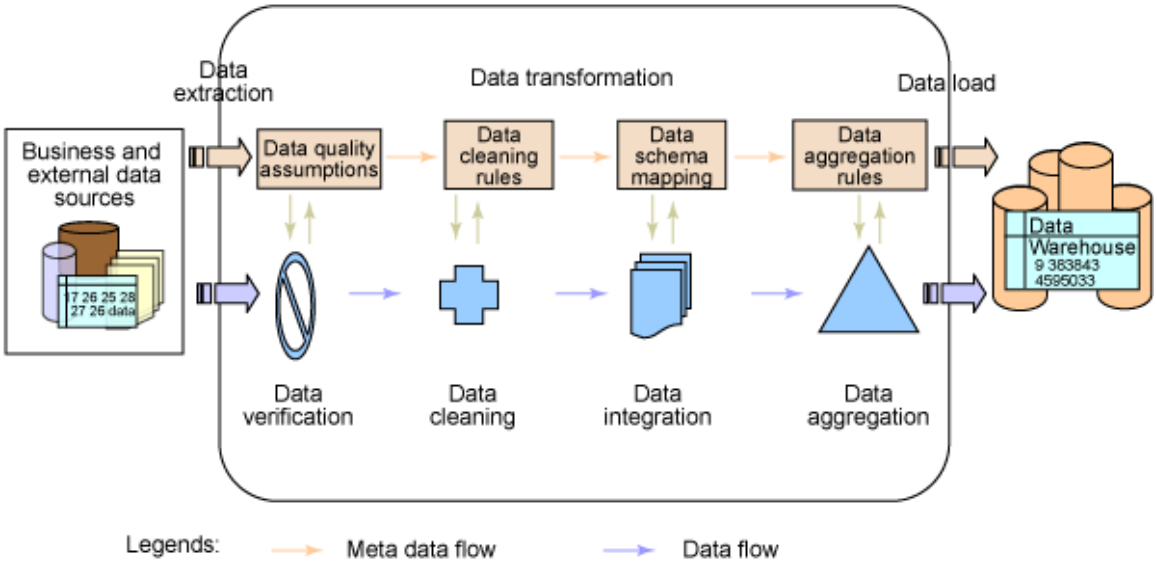


Figura 2.- Procés ETL

Revisant els fitxers font, s'observen diverses diferències entre els fitxers aportats per les CCAA amb les dades tributàries. Tot i tenir una estructura semblant, no tots els fitxers donen la mateixa informació. Cal tenir en compte que:

- Algunes comunitats donen informació tributària de l'any 2002 al 2009, altres del 2003 al 2009 i d'altres del 2004 al 2009.
- Alguns inclouen la informació dels totals de totes les dades dels diferents tipus de perceptors i d'altres els totals de totes les dades dels diferents tipus de prestacions.
- Alguns inclouen dades de retribucions mitges i tipus mig de retenció, altres fitxers contenen aquestes dades en valors absoluts i d'altres proporciones els dos valors.
- El fitxer de Ceuta i Melilla separa les dades per les dues ciutats autònomes, com si fossin separades, però totes les dades dintre del mateix fitxer.
- El registres que no contenen dades vàlides tenen ".." dintre del camp.
- Totes les dades, tot i ser numèriques, vénen donades com un string entre cometes. Per exemple: "15.75".
- Hi ha dades tributàries de l'any 2002 al 2009, dades de població del 2005 al 2011 i dades de població activa del 2005 al 2010.
- A més, degut a que la informació s'ha extret de diferents sistemes, hi ha tres formats de fitxer CSV diferents: camps separats per punt i coma, per coma o tabulador.

En els fitxers de població també hi ha aspectes a tenir en compte:

- Els fitxers amb les dades de població inclouen el País Basc i Navarra. Aquestes dades no són necessàries perquè no es tenen les dades tributàries.
- Es demana que, per a que les dades siguin més realistes, el nombre d'habitants es calculi com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent.

A l'apartat 3.5 s'explica la implementació dels processos ETL.

2.3. Volum dels elements d'anàlisis

Per veure la quantitat de dades amb les que es treballa s'importen els fitxers CSV a Excel. Separats per tipus de fitxer es té:

- *Población por CCAA xxxx*: 1 columna amb els noms de les CCAA, 1 columna de dades per cada sexe i un registre amb la quantitat de població per cadascun. Això per cada CA i els totals. Fan un total de 20 registres per cada sexe i pels noms de les CCAA. Hi han 7 fitxers (de l'any 2005 al 2011), total de 420 registres.

- *Porc Población Activa*: 6 columnes (de l'any 2005 al 2010) amb un registre per cada any i CA i els totals. A més, hi ha una columna amb els noms de les CCAA. Fan un total de 19 registres per 7 columnes, total de 133 registres.

- *Tributación XXXXXXX*: 7 columnes de dades dels diferents tipus de perceptors, amb 15 registres per cada columna, més una columna amb els tipus de perceptors. Anys del 2002 al 2009, fan un total de 50 columnes. Total 750 registres a cada un del 16 fitxers de les diferents CCAA = 12000 registres.

Com es dedueix, són masses registres com per poder treballar d'una forma còmoda amb fulls de càlcul.

2.4. Objectius i informes a implementar

L'objectiu general del projecte consisteix en la creació d'un magatzem de dades, que reculli i realitzi les tasques d'explotació de les dades de prestacions d'una manera fàcil i el més automatitzada possible. A més, ha de facilitar la generació d'informes per ajudar a la posterior presa de decisions. Amb els resultats obtinguts, L'OADT podrà determinar la sostenibilitat del sistema de prestacions socials i conèixer l'evolució d'aquestes prestacions.

Es té en compte que la part de càrrega de dades ha de ser un procés automàtic i que, a més, cal que transformi el contingut d'alguns camps per solucionar els problemes que s'han comentat al punt 2.2. Aquest últim punt és un tema important en el projecte, i és el de prendre decisions sobre temes no descrits o amb comportaments diferents als que s'esperen. Idealment, s'hauria de buscar la correspondència amb les regles definides.

Un cop es té el magatzem de dades operatiu amb les dades carregades, es poden realitzar les consultes requerides pel client:

- Total retribució
- Total retenció
- Retribució mitjana
- % de retenció mig (import retingut respecte del total percebut)
- Número de retribucions mig (si una persona rep pensió y salari comptarà com dos)
- % de població per segment (activa, pensionistes, aturats, …)
- Nombre de treballadors/Nombre de perceptors no actius (Determina la sostenibilitat del sistema)
- Número de treballadors/Habitants
- Número de treballadors/Número de persones actives
- Salaris totals/Total Prestacions

Aquesta informació es proporciona dintre d'una temporalitat a nivell d'any i es pot consultar de forma agregada per CA, tipus de perceptor i tipus de retribució.

Es demana la possibilitat d'obtenir la projecció dels indicadors "Num Treb/Num Perceptors no actius" i "Num Treb/Població" per determinar el punt de col·lapse, és a dir, el moment en que dos treballadors mantinguin un perceptor, o s'iguali el nombre de treballadors i població perceptora (no actius + perceptors de prestacions socials).

2.5. Diagrama de casos d'ús

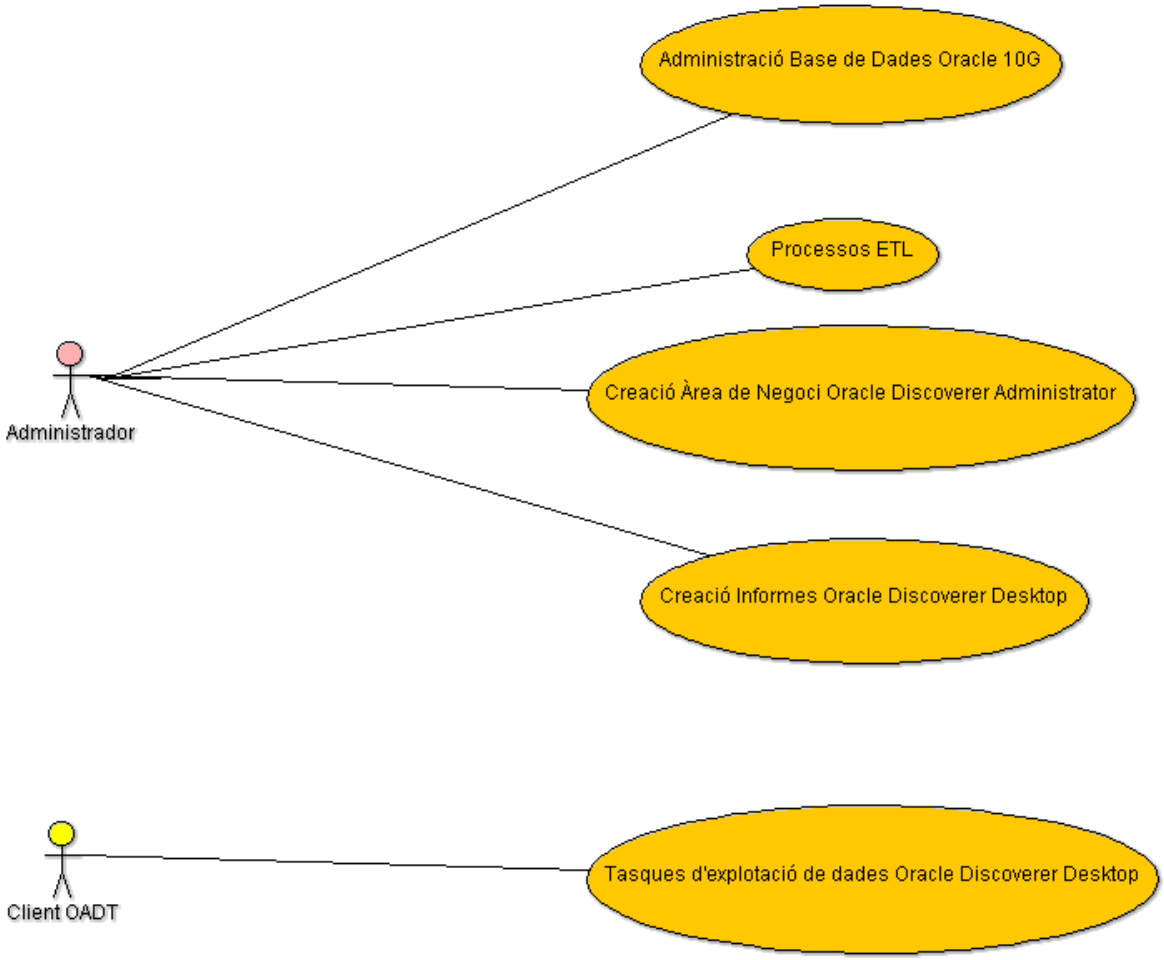


Figura 3.- Casos d'ús

L'actor Administrador és l'encarregat de la creació del magatzem de dades. Gestiona la base de dades, realitza la càrrega i manipulació dels fitxers font, crea l'àrea de negoci i implementa els informes que donen resposta a les consultes que demana el client.

L'actor Client OADT és l'usuari encarregat de l'explotació de dades sobre el model creat.

2.6. Model conceptual

Per les dades de què es disposa i el tipus d'informes que es volen obtenir, es creu que el més apropiat és utilitzar una base de dades OLAP (de l'anglès, *On-Line Analytical Processing*). És una solució utilitzada en el camp de la Intel·ligència empresarial (o Business Intelligence) que té l'objectiu d'agilitzar la consulta de grans quantitats de dades.

Dintre dels sistemes OLAP es troba el concepte de cub OLAP (cub multidimensional o hipercub). Es composa de "Fets" numèrics anomenats "Mesures" que es classifiquen per dimensions.

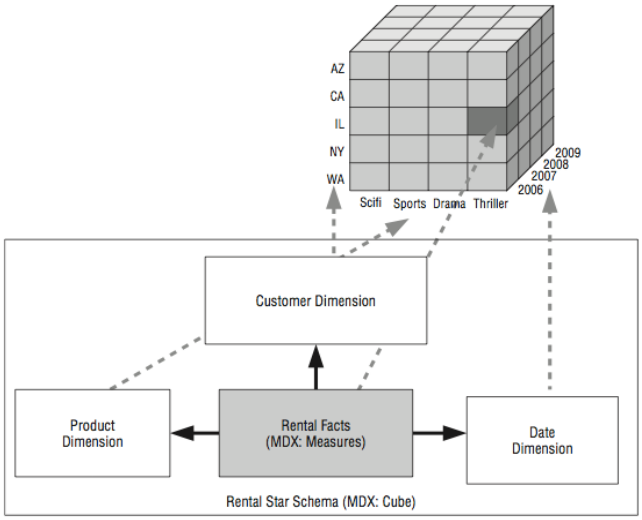


Figura 4.- Exemple de Cub OLAP

El cub OLAP es crea a partir d'un esquema en "Estrella" o "Floc de neu". El primer és més senzill i es composa d'una taula de Fets, que conté les dades per l'anàlisi, envoltada de taules de dimensions. En canvi, l'estructura de "Floc de neu" es fa servir quan alguna de les dimensions s'implementa amb més d'una taula de dades.

Per la implementació del present projecte es decideix utilitzar l'esquema en "Estrella", ja que cap dimensió necessita més d'una taula. Es pot veure a continuació que el disseny es composa de dues estrelles, ja que es tenen dos Fets.

Per la realització del disseny conceptual cal seguir els passos que es detallen a continuació. Amb les dades que queden determinades (fets, granularitat, dimensions, atributs, mesures, etc) es pot obtenir l'estrella doble que modelitzarà el disseny.

2.6.1. Triar els fets

El fet és la part central del model que es vol definir. Aquesta part central és una taula que s'anomena taula de fet i conté els valors de les mesures de negoci. Cada mesura es pren mitjançant la intersecció de les dimensions que la defineixen. Aquestes dimensions estan reflectides a les seves corresponents taules de dimensions, que envolten la taula de fet i estan relacionades amb ella.

Segons els requeriments del projecte, es decideix tenir dos fets diferenciats. Són els següents:

- Prestació: són les prestacions que tenen assignades els ciutadans i que acompleixen certs requisits com tipus de perceptor, tipus de retribució, la CA i any als que pertanyen. La taula de fets inclou el número de persones que reben aquesta prestació, de quina quantitat és la retribució i la seva retenció. Aquest és el fet principal.

- Població: són les dades d'habitants i població activa per cada CA i any d'estudi.

El motiu de tenir un fet de població és perquè les dades de població i població activa són relatives a una CA i un any (que són dues dimensions comunes als dos fets del model de dades). Amb això es té un model amb dues estrelles relacionades.

2.6.2. Granularitat

Una característica important que defineix les taules de fets és el nivell de granularitat de les dades que s'emmagatzemen. La granularitat és el nivell de detall de les dades, és a dir, la granularitat de la taula de fet representa el nivell més atòmic pel qual es defineixen les dades.
La granularitat afecta a la cardinalitat, tant de les dimensions com de la taula de fet. A major granularitat (gra més prim) major serà el número de registres final de la taula de fet. Cal tenir en compte que a major detall més càrrega de treball pel sistema, el que comporta més lentitud en l'execució dels processos. Per contra, menor granularitat representa perdre informació. Així que cal tenir molt clar quin grau de detall es necessita per no fer treballar el sistema innecessàriament.

Al present projecte es tenen prestacions anuals per cada CA, tipus de perceptor i tipus de retribució per un fet i, dades d'habitants i població activa per CA i any pel segon fet. Sembla ja una granularitat adequada que no sobrecarrega el sistema i que aporta el nivell de detall suficient per realitzar les posteriors consultes i informes.

2.6.3. Triar les dimensions

La construcció de cubs OLAP requereix de taules de fets i diverses taules de dimensions. Aquestes últimes determinen els paràmetres (dimensions) dels que depenen els fets enregistrats a les taules de fets. O sigui, que aquelles dades que es guarden a les taules de dimensions són les que ajuden a definir el fet pròpiament dit.
Les taules de dimensions contenen atributs que s'utilitzen per restringir i agrupar les dades emmagatzemades a una taula de fet quan es realitzen consultes sobre aquella dada.
Hi ha atributs que són descriptors, que permeten fer seleccions basades en els seus valors, i els que permeten fer agrupacions.
Segons els requisits del client i les dades aportades, es creu necessàries les següents dimensions:

- Dimensio_TipusPerceptor. Indica el tipus de perceptor, que pot ser assalariat, pensionista, etc.
- Dimensio_TipusRetribucio. Indica el tipus de retribució, que pot ser salari, pensió, etc.
- Dimensio_Temps. És la dimensió temporal. Només interessa a nivell d'any.
- Dimensio_ComunitatAutonoma. Indica informació de la CA com nom, població activa, etc.

El fet prestació està relacionat amb les quatre dimensions i el fet població es relaciona únicament amb les dimensions Temps i ComunitatAutonoma.

2.6.4. Triar els atributs de les dimensions

Dintre de cada dimensió es troben els diferents atributs que hi formen part i que seran útils per a triar i descriure l'espai d'anàlisi. De cada dimensió hi ha els següents:

- Dimensio_TipusPerceptor. Atribut Tipus, que indica quin tipus de perceptor és (Assalariat, Aturat, Pensionista, etc) i un atribut que indica el número de prestacions que rep aquest perceptor. Aquest últim atribut és útil per poder calcular el número de retribucions mig, ja que si una persona rep més d'una prestació alhora comptaria com més d'un.

- Dimensio_TipusRetribucio. Atribut Tipus, que indica quin tipus de retribució és (Salari, Pensió o Atur). Tot i que aquest atribut podria estar al fet prestació, cal posar-lo en una dimensió independent per posteriorment realitzar les consultes agregades que requereix el client.

- Dimensio_Temps. Les retribucions, retencions i habitants, que determinen els fets, venen determinades per l'Any al que corresponen les dades. Tota la informació demanada pel client es proporciona dintre d'una temporalitat a nivell d'any.

- Dimensio_ComunitatAutonoma. Guarda l' atribut Nom de la CA.

2.6.5. Les mesures dels fets

- Fet Prestació. Les mesures que es tenen en compte són les dades referents a les prestacions per any i CA. Les bàsiques són el número de persones que reben una prestació, l'import de la retribució per aquesta prestació i quina retenció se li aplica a la mateixa. A partir d'aquestes dades, el client pot determinar la sostenibilitat del sistema de prestacions socials i conèixer l'evolució d'aquestes prestacions.

- Fet Població. Les mesures que es tenen en compte són el número d'habitants, separats per homes i dones, i el percentatge de població activa. Aquestes dades són independents per cada CA i any d'estudi.

2.6.6. Definir les cel·les

Cal definir quines cel·les resultants de l'anàlisi són interessants pel model, així que cal que es guardin.
Es té una única cel·la pel fet prestació, que és la que identifica una prestació en concret, pertanyent a un tipus de perceptor i que té un tipus de retribució associada. Pel fet població es té també una única cel·la que representa les dades de població i població activa. En ambdós casos això relatiu a una CA i un any en concret.

2.6.7. Explicitar les restriccions d'integritat

És necessari definir les bases i establir quines són les restriccions d'integritat. El conjunt inicial de dimensions, que es troben després de definir el gràNul, dóna lloc a una base. La base ajuda a

saber quines són les dimensions realment necessàries perquè una cel·la sigui única. Segons el model de dades, hi ha 4 dimensions: Temps, ComunitatAutonoma, TipusPerceptor i TipusRetribucio. Es sap que una prestació en concret només existeix una vegada per una CA, un tipus de perceptor, un tipus de retribució en un any. Es pot veure que una cel·la realment està determinada per les quatre dimensions, tot i que és molt difícil que es repeteixen els valors exactament entre diferents prestacions. El mateix passa amb les dades de població, unes dades en concret venen determinades per una CA i un any. Aquesta cel·la està determinada per aquestes dues dimensions.

2.6.8. Estudiar la viabilitat

Un cop realitzat el disseny conceptual, cal veure si es pot implementar o no. És necessari tenir una estimació de la memòria que faran servir les dades al sistema.

Amb els fitxers aportats pel client amb les dades tributàries es tenen com a màxim 2352 instàncies del fet prestació, resultat de multiplicar els 7 tipus de perceptors per 3 (que és el màxim de retribucions diferents que pot rebre) i això multiplicar-lo per 7 anys i per 16 fitxers. De cada fet es guarda un apuntador que és un Integer a cadascuna de les quatre dimensions, el número de persones que rep una prestació és un Integer, les retribucions i retencions són Reals. Cada instància és aproximadament de 4 bytes pels Integers i reals, el que fan 5 Integers x 4 bytes + 2 Reals x 4 bytes = 28 bytes.
Totes les instàncies ocupen aproximadament 28 bytes x 2352 = 65856 bytes, el que és el mateix que 64 Kbytes.

Pel que fa a les dades de població, es tenen un màxim de 119 instàncies del fet població, resultat de multiplicar el número de CCAA pels 7 anys d'estudi (del 2005 al 2011). De cada fet es guarda un apuntador que és un Integer a cadascuna de les dues dimensions, el número d'homes i dones són Integers i el percentatge de població activa és un real. Cada instància és aproximadament de 4 Integers x 4 bytes + 1 Reals x 4 bytes = 20 bytes.
Totes les instàncies ocupen aproximadament 20 bytes x 119 = 2380 bytes, el que és el mateix que 2,3 Kbytes.

Es dedueix que la despesa de memòria és bastant reduïda per assumir-la amb qualsevol equip informàtic actual, sense veure's deteriorat el rendiment del mateix ni a l'hora de realitzar cap operació sobre la base de dades, incloses les consultes.

S'ha de tenir en compte que aquesta és una estimació, ja que les dimensions també ocupen memòria però poca comparada amb els Fets. A més, no tots els perceptors reben els tres tipus de retribucions i s'ha considerat que si.

2.6.9. Resultat

El model conceptual resultant és el següent:

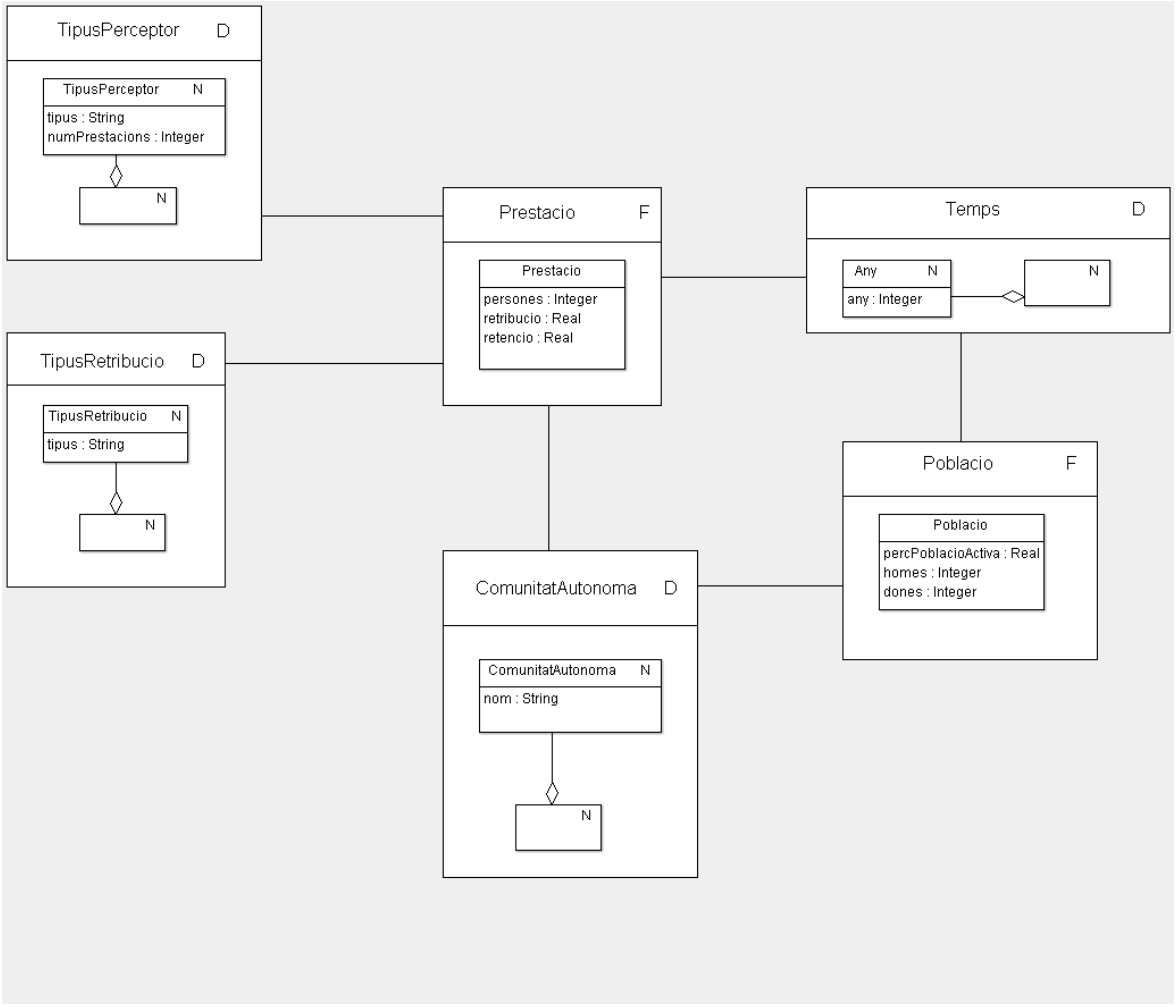


Figura 5.- Diagrama del model conceptual

3. Disseny

3.1. Arquitectura del maquinari

A la imatge següent es pot veure l'arquitectura del maquinari proposada per la posada en marxa del magatzem de dades i la posterior explotació.

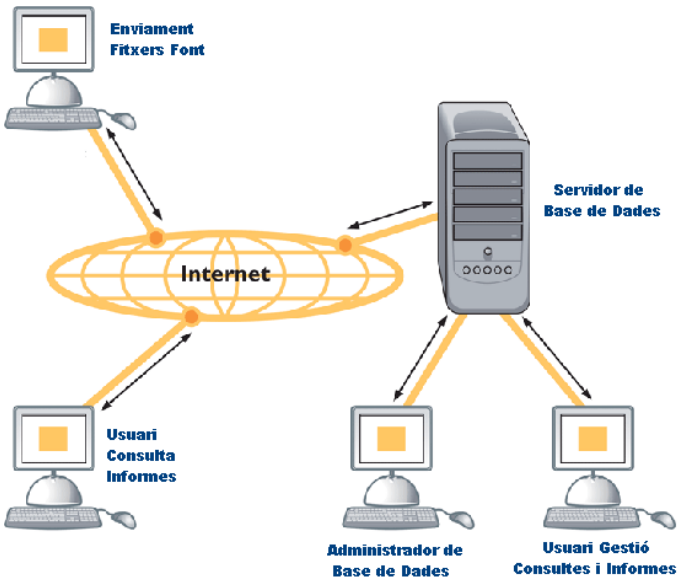


Figura 6.- Arquitectura del maquinari

Consta d'un servidor central on es troba implementat el Data Warehouse, amb el corresponent sistema gestor de base de dades. Allà es realitzen els processos ETL quan arriben noves dades i s'executen les consultes i informes. El client envia, via Internet, els fitxers que serveixen com origen de dades i realitza les tasques d'extracció de dades en alguna estació de treball. Per altra banda, els administradors i usuaris encarregats de la gestió de les consultes i informes gestionen el magatzem de dades remotament, dintre de la xarxa d'àrea local de l'empresa.

3.2. Arquitectura del programari

A la imatge següent es pot veure l'arquitectura del programari proposada per la posada en marxa del magatzem de dades i la posterior explotació.

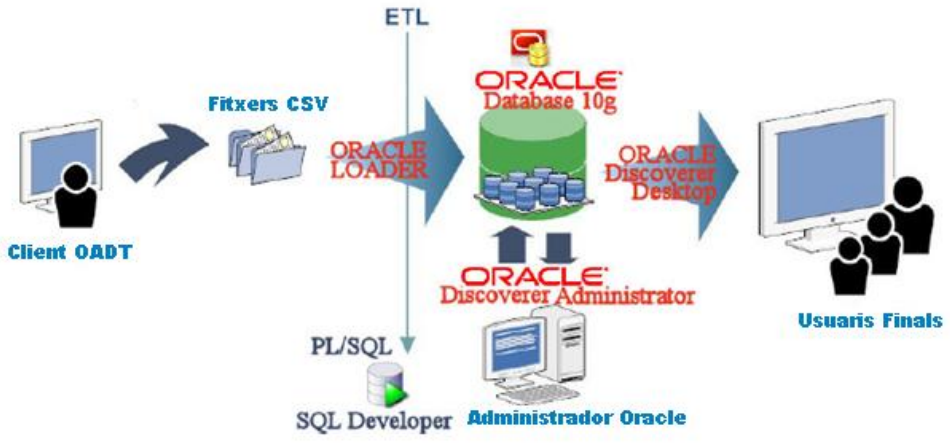


Figura 7.- Arquitectura del programari

Els fitxers aportats pel client es carreguen amb sqlloader a les taules d'Oracle Database 10G. La creació de taules, els processos ETL i la càrrega a les taules definitives es realitzen en SQL i PL/SQL mitjançant SQL Developer. La creació de l'àrea de negoci es realitza amb Oracle Discoverer Administrator i les consultes i informes amb Oracle Discoverer Desktop.

3.3. Disseny lògic

Completat el disseny conceptual es realitza el disseny lògic.

Es necessita una taula per cada dimensió i una altra taula per cada fet. Les taules de fets contenen, a part de les mesures indicades al disseny conceptual, claus foranes a les taules de dimensions. La clau primària de la taula de fet prestació és el conjunt de claus foranes, ja que cada prestació està identificada per un tipus de perceptor, un tipus de retribució, una CA i un any en concret.
La clau primària de la taula de fet població està composta per les claus foranes a les taules de dimensió Temps i ComunitatAutonoma, ja que fan que cada fet sigui únic.

Les taules de les dimensions contenen, a part dels atributs indicats anteriorment, les corresponents claus primàries que permeten enllaçar les taules de dimensions amb les de fets.

Una manera de reduir la grandària de les taules del Fets, que ocupen la major part de l'espai, és definir substituts de la clau primària a les taules de dimensió. Així, a la dimensió TipusPerceptor, on es té un atribut identificador únic que és "tipus" i que és de tipus String, s'utilitza un nou atribut que ocupa menys espai al ser de tipus Integer. Es fa el mateix amb la resta de taules de dimensions. D'aquesta forma es redueix la grandària de les columnes que formen la clau primària de les taules de Dimensió i, consegüentment, la grandària de la clau primària de les taules dels Fets. A més, es prevenen possibles problemes davant de canvis en els atributs identificadors del sistema operacional, ja que és probable que en algun moment hi hagin recodificacions dels identificadors que farien que s'haguessin de recodificar totes les dades. Com que ara l'identificador és independent de l'origen de les dades, qualsevol canvi no l'afecta.

Segons això, s'obté el següent model lògic:

Prestacio(idAny, idTipusPercep, idTipusRetrib, idComAut, persones, retribucio, retencio)
Poblacio(idAny, idComAut, percPoblacioActiva, homes, dones)
TipusPerceptor(RowID, tipus, numPrestacions)
TipusRetribucio(RowID, tipus)
ComunitatAutonoma(RowID, nom)
Temps(RowID, any)

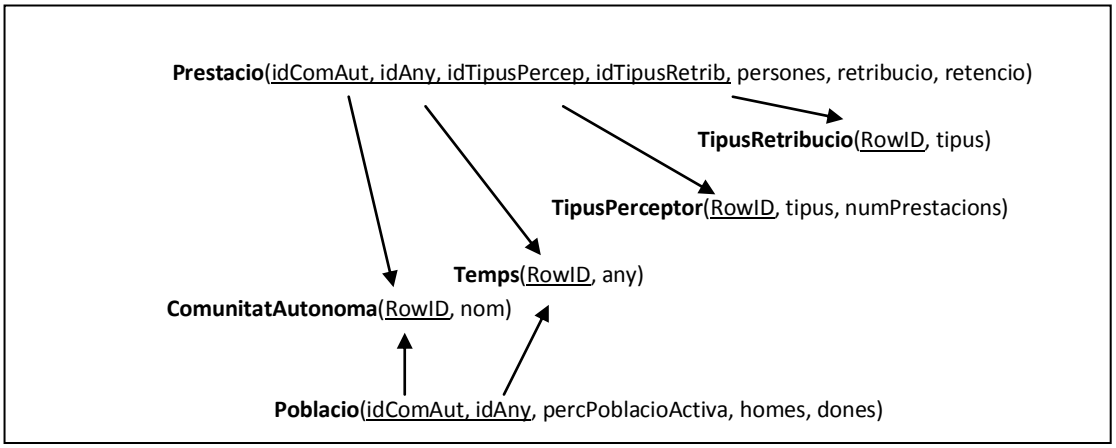
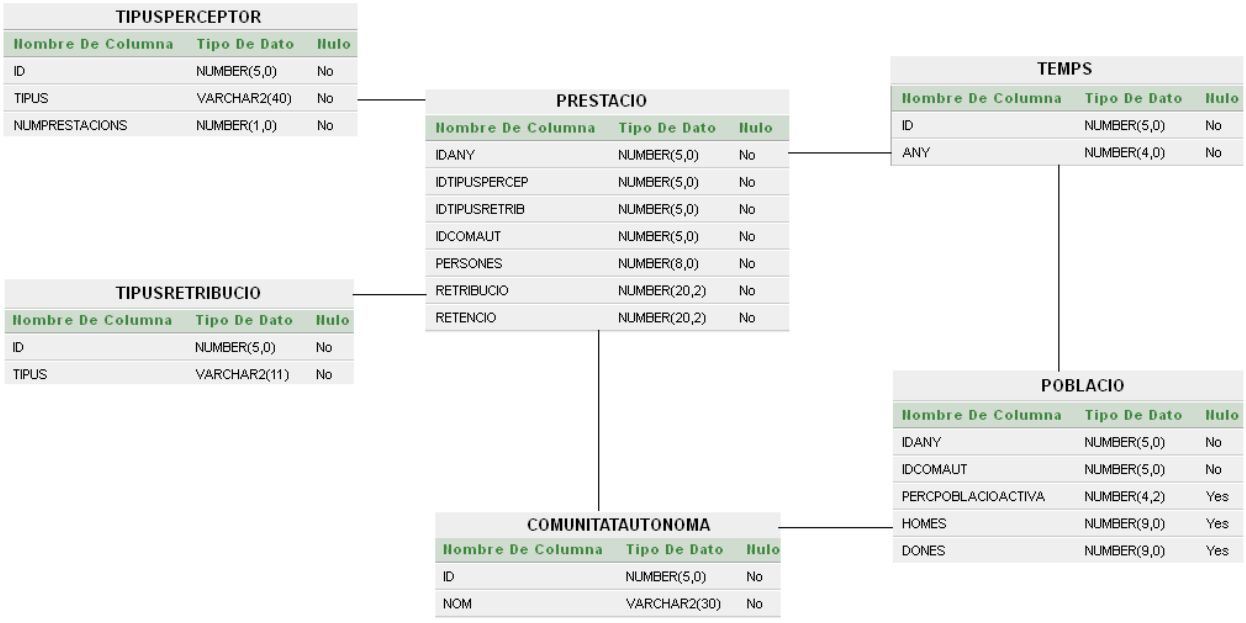


Figura 8.- Disseny de la BD

3.4. Disseny físic

3.4.1. Disseny de les taules i relacions

Definit el disseny lògic, amb les corresponents taules i relacions, ja es pot realitzar el disseny físic. Aquest disseny inclou les taules necessàries pel projecte i que són gestionades per un motor de bases de dades. Es fa servir SqlDeveloper per la creació de les taules. Són les següents:



TIPUSPERCEPTOR

Nombre De Columna	Tipo De Dato	Nulo
ID	NUMBER(5,0)	No
TIPUS	VARCHAR2(40)	No
NUMPRESTACIONS	NUMBER(1,0)	No

PRESTACIO

Nombre De Columna	Tipo De Dato	Nulo
IDANY	NUMBER(5,0)	No
IDTIPUSPERCEP	NUMBER(5,0)	No
IDTIPUSRETRIB	NUMBER(5,0)	No
IDCOMAUT	NUMBER(5,0)	No
PERSONES	NUMBER(8,0)	No
RETRIBUCIO	NUMBER(20,2)	No
RETENCIO	NUMBER(20,2)	No

TEMPS

Nombre De Columna	Tipo De Dato	Nulo
ID	NUMBER(5,0)	No
ANY	NUMBER(4,0)	No

TIPUSRETRIBUCIO

Nombre De Columna	Tipo De Dato	Nulo
ID	NUMBER(5,0)	No
TIPUS	VARCHAR2(11)	No

POBLACIO

Nombre De Columna	Tipo De Dato	Nulo
IDANY	NUMBER(5,0)	No
IDCOMAUT	NUMBER(5,0)	No
PERCPOBLACIOACTIVA	NUMBER(4,2)	Yes
HOMES	NUMBER(9,0)	Yes
DONES	NUMBER(9,0)	Yes

COMUNITATAUTONOMA

Nombre De Columna	Tipo De Dato	Nulo
ID	NUMBER(5,0)	No
NOM	VARCHAR2(30)	No

Figura 9.- Esquema del disseny de la BD

A continuació s'explica cada taula i el disseny de les mateixes, amb una explicació de cada camp.

PRESTACIO

Taula que guarda les instàncies del fet de l'estrella principal, que són les dades de les prestacions.

- IDANY: conté l'identificador de l'any al que pertany la prestació. Forma part de la clau primària de la taula i és clau forana a la dimensió Temps, per tant, no pot ser NULL.
- IDTIPUSPERCEP: conté l'identificador del tipus de perceptor al que pertany la prestació. Forma part de la clau primària de la taula i és clau forana a la dimensió Tipusperceptor, per tant, no pot ser NULL.
- IDTIPUSRETRIB: conté l'identificador del tipus de retribució al que pertany la prestació. Forma part de la clau primària de la taula i és clau forana a la dimensió Tipusretribucio, per tant, no pot ser NULL.
- IDCOMAUT: conté l'identificador de la CA a la que pertany la prestació. Forma part de la clau primària de la taula i és clau forana a la dimensió Comunitatautonoma, per tant, no pot ser NULL.
- PERSONES: conté el valor que indica el número de persones que reben la prestació. Es considera que el seu valor no pot ser NULL.
- RETRIBUCIO: conté l'import total de retribució de les instàncies del fet. Es considera que el seu valor no pot ser NULL.
- RETENCIO: conté l'import de la retenció aplicada a la retribució de les instàncies del fet. Es considera que el seu valor no pot ser NULL.

COMUNITATAUTONOMA

Taula de dimensió que guarda les CCAA objecte d'estudi.

- ID: conté l'identificador de la CA. És la clau primària, per tant, no pot ser NULL.
- NOM: conté el nom de la CA. Es considera que el seu valor no pot ser NULL.

TEMPS

Taula de dimensió que guarda els anys objecte d'estudi. És la taula de dimensió temporal.

- ID: conté l'identificador de l'any. És la clau primària, per tant, no pot ser NULL.
- ANY: conté el número d'any. Es considera que el seu valor no pot ser NULL.

TIPUSPERCEPTOR

Taula de dimensió que guarda els tipus de perceptor que existeixen (Assalariat, Aturat, etc).

- ID: conté l'identificador del tipus de perceptor. És la clau primària, per tant, no pot ser NULL.

- TIPUS: conté el tipus de perceptor. Es considera que el seu valor no pot ser NULL.

- NUMPRESTACIONS: conté el número de prestacions diferents que rep. Es considera que el seu valor no pot ser NULL.

TIPUSRETRIBUCIO

Taula de dimensió que guarda els tipus de retribucions que existeixen (Salari, Atur, etc).

- ID: conté l'identificador del tipus de retribució. És la clau primària, per tant, no pot ser NULL.

- TIPUS: conté el tipus de retribució. Es considera que el seu valor no pot ser NULL.

POBLACIO

Taula que guarda les instàncies del fet de l'estrella secundària. És la informació de població per any i CA.

- IDANY: conté l'identificador de l'any al que pertanyen les dades de població. Forma part de la clau primària de la taula i és clau forana a la dimensió Temps, per tant, no pot ser NULL.

- IDCOMAUT: conté l'identificador de la CA a la que pertanyen les dades de població. Forma part de la clau primària de la taula i és clau forana a la dimensió Comunitatautonoma, per tant, no pot ser NULL.

- PERCPOBLACIOACTIVA: guarda el percentatge de població activa.

- HOMES: guarda el número d'homes.

- DONES: guarda el número de dones.

A continuació es mostren els scripts de creació de taules:

```
CREATE TABLE "POBLACIO"
  (         "IDANY" NUMBER(5,0) NOT NULL ENABLE,
            "IDCOMAUT" NUMBER(5,0) NOT NULL ENABLE,
            "PERCPOBLACIOACTIVA" NUMBER(4,2),
            "HOMES" NUMBER(9,0),
            "DONES" NUMBER(9,0),
             CONSTRAINT "POBLACIO_PK" PRIMARY KEY ("IDANY", "IDCOMAUT") ENABLE,
             CONSTRAINT "ANY_FK1" FOREIGN KEY ("IDANY")
             REFERENCES "TEMPS" ("ID") ENABLE,
             CONSTRAINT "COMAUT_FK1" FOREIGN KEY ("IDCOMAUT")
```

```
        REFERENCES "COMUNITATAUTONOMA" ("ID") ENABLE
 ) ;

CREATE TABLE "PRESTACIO"
 (      "IDANY" NUMBER(5,0) NOT NULL ENABLE,
        "IDTIPUSPERCEP" NUMBER(5,0) NOT NULL ENABLE,
        "IDTIPUSRETRIB" NUMBER(5,0) NOT NULL ENABLE,
        "IDCOMAUT" NUMBER(5,0) NOT NULL ENABLE,
        "PERSONES" NUMBER(8,0) NOT NULL ENABLE,
        "RETRIBUCIO" NUMBER(20,2) NOT NULL ENABLE,
        "RETENCIO" NUMBER(20,2) NOT NULL ENABLE,
        CONSTRAINT "PRESTACIO_PK" PRIMARY KEY ("IDANY", "IDTIPUSPERCEP",  "IDTIPUSRETRIB", "IDCOMAUT")
       ENABLE,
        CONSTRAINT "ANY_FK2" FOREIGN KEY ("IDANY")
        REFERENCES "TEMPS" ("ID") ENABLE,
        CONSTRAINT "COMAUT_FK2" FOREIGN KEY ("IDCOMAUT")
        REFERENCES "COMUNITATAUTONOMA" ("ID") ENABLE,
        CONSTRAINT "TIPUSPERCEP_FK1" FOREIGN KEY ("IDTIPUSPERCEP")
        REFERENCES "TIPUSPERCEPTOR" ("ID") ENABLE,
        CONSTRAINT "TIPUSRETRIB_FK1" FOREIGN KEY ("IDTIPUSRETRIB")
        REFERENCES "TIPUSRETRIBUCIO" ("ID") ENABLE
 ) ;

 CREATE TABLE "TEMPS"
 (      "ID" NUMBER(5,0) NOT NULL ENABLE,
        "ANY" NUMBER(4,0) NOT NULL ENABLE,
        CONSTRAINT "TEMPS_PK" PRIMARY KEY ("ID") ENABLE
 ) ;

 CREATE TABLE "TIPUSPERCEPTOR"
 (      "ID" NUMBER(5,0) NOT NULL ENABLE,
        "TIPUS" VARCHAR2(40 BYTE) NOT NULL ENABLE,
        "NUMPRESTACIONS" NUMBER(1,0) NOT NULL ENABLE,
        CONSTRAINT "TIPUSPERCEPTOR_PK" PRIMARY KEY ("ID") ENABLE
 ) ;

 CREATE TABLE "TIPUSRETRIBUCIO"
 (      "ID" NUMBER(5,0) NOT NULL ENABLE,
        "TIPUS" VARCHAR2(11 BYTE) NOT NULL ENABLE,
        CONSTRAINT "TIPUSRETRIBUCIO_PK" PRIMARY KEY ("ID") ENABLE
 ) ;

 CREATE TABLE "COMUNITATAUTONOMA"
 (      "ID" NUMBER(5,0) NOT NULL ENABLE,
        "NOM" VARCHAR2(30 BYTE) NOT NULL ENABLE,
        CONSTRAINT "COMUNITATAUTONOMA_PK" PRIMARY KEY ("ID") ENABLE
 ) ;
```

A part de les taules indicades, es fan servir unes taules temporals que ajuden a fer la càrrega de les dades a les taules definitives:

```
CREATE TABLE "TRIBUTACIO"
 (      "DADES" VARCHAR2(30),
        "ASAL2002" NUMBER(20,2),
        "PENSIO2002" NUMBER(20,2),
        "DESEMPL2002" NUMBER(20,2),
```

```
        "ASALPENSIO2002" NUMBER(20,2),
        "ASALDESEMPL2002" NUMBER(20,2),
        "PENSIODESEMPL2002" NUMBER(20,2),
        "ASALPENSIODESEMPL2002" NUMBER(20,2),
        ..............
        "ASAL2009" NUMBER(20,2),
        "PENSIO2009" NUMBER(20,2),
        "DESEMPL2009" NUMBER(20,2),
        "ASALPENSIO2009" NUMBER(20,2),
        "ASALDESEMPL2009" NUMBER(20,2),
        "PENSIODESEMPL2009" NUMBER(20,2),
        "ASALPENSIODESEMPL2009" NUMBER(20,2)
) ;
```

Aquesta taula serveix per fer la càrrega de les dades tributàries tal com venen donades als fitxers CSV. A partir d'aquesta taula s'extreuen les dades per distingir les prestacions.

```
CREATE TABLE "TEMPPREST"
(       "DADES" VARCHAR2(30),
        "ASAL2002" NUMBER(20,2),
        "PENSIO2002" NUMBER(20,2),
        "DESEMPL2002" NUMBER(20,2),
        "ASALPENSIO2002" NUMBER(20,2),
        "ASALDESEMPL2002" NUMBER(20,2),
        "PENSIODESEMPL2002" NUMBER(20,2),
        "ASALPENSIODESEMPL2002" NUMBER(20,2),
        ..............
        "ASAL2009" NUMBER(20,2),
        "PENSIO2009" NUMBER(20,2),
        "DESEMPL2009" NUMBER(20,2),
        "ASALPENSIO2009" NUMBER(20,2),
        "ASALDESEMPL2009" NUMBER(20,2),
        "PENSIODESEMPL2009" NUMBER(20,2),
        "ASALPENSIODESEMPL2009" NUMBER(20,2)
);
```

Aquesta taula serveix per anar guardant les dades d'una prestació en concret. Aquestes s'incorporen a la taula definitiva quan es tenen totes les dades de la mateixa instància del fet.

```
CREATE TABLE "TMPCOMAUTON"
(       "COMAUT" VARCHAR2(30 BYTE),
        "HOME" NUMBER(9,0),
        "DONA" NUMBER(9,0)   ) ;
```

Aquesta taula serveix per guardar totes les dades de tots els fitxers amb informació de població tal com venen dels fitxers CSV. A partir d'aquesta taula s'extreuen les dades per omplir la taula definitiva amb informació de cada CA.

```
CREATE TABLE "TMPCOMAUTONPERC"
(       "COMAUT" VARCHAR2(30 BYTE),
        "2010" NUMBER(4,2),
        "2009" NUMBER(4,2),
        "2008" NUMBER(4,2),
        "2007" NUMBER(4,2),
        "2006" NUMBER(4,2),
```

```
        "2005" NUMBER(4,2)
) ;
```

Aquesta taula serveix per guardar les dades del fitxer amb informació de població activa. A partir d'aquesta taula s'extreuen les dades per omplir la taula definitiva amb informació de cada CA.

El script de creació de taules es pot consultar al fitxer *creataules.sql*. En aquest script també es troben les sentències SQL per realitzar la càrrega de les dades a les taules de dimensió. Tots els fitxer utilitzats pel desenvolupament del projecte es troben al disc de la màquina virtual (UOCPRACT.vdi, path E:\fitxers).

Tot el disseny de la base de dades es realitza amb SqlDeveloper. Prèviament, es crea l'usuari USERTFC amb l'aplicació Oracle Database Express Edition.

Figura 10.- Connexió a Oracle Express Edition amb usuari creat

Les dades de connexió són les següents:

Usuari: **USERTFC**
Paraula de pas: **uoc**

3.4.2. Índexs

Per la creació dels índexs es considera el volum de dades a tractar i el disseny de la base de dades. Els índexs més importants són els creats sobre les claus primàries i foranes de totes les taules. Aquests índexs es creen de manera implícita pel propi Oracle, com a conseqüència de les restriccions PRIMARY KEY (es crea un índex únic sobre els camps clau) i FOREIGN KEY (es crea un índex amb possibilitat de repetir valors, és un índex amb duplicats). No és necessari crear índexs més específics per les consultes indicades, ja que el temps d'execució de les mateixes és baix.

3.5. Procés d'extracció, transformació i càrrega de dades

Creades les taules ja es pot realitzar la càrrega de les dades proporcionades per l'OADT.
Com ja s'ha comentat, es disposa d'un fitxer en format CSV per cada CA amb la informació tributària dels diferents anys. Per automatitzar la carrega s'utilitza l'aplicació sqlloader.
Sqlloader treballa amb un fitxer de control on s'especifica el fitxer de dades a tractar, delimitadors de camps, columnes on inserir les dades, etc.

El contingut bàsic dels fitxers de control és:

```
LOAD DATA
INFILE c:\ruta_completa\archivo_data.csv
BADFILE c:\ruta_completa\bad_file.bad
INTO TABLE <esquema>.<tabla>
APPEND
FIELDS TERMINATED BY ','
TRAILING NULLCOLS
(campo1,
campo2,
campo3)
```

És necessari crear un fitxer de control per cada CA, ja que cada una d'elles aporta el fitxer amb un format propi. Els delimitadors de camps no són els mateixos (uns tenen la coma, altres tabulador i d'altres punt i coma), no tots tenen dades tributàries del 2002 al 2009 (el que vol dir que hi han columnes buides per aquests anys i que s'han d'indicar al fitxer de control) i alguns inclouen columnes amb dades totals.

S'aprofita l'opció d'utilitzar més d'un delimitador per indicar que les cometes dobles és un terminador de camp, d'aquesta manera es poden tractar el valors directament com numèrics, estalviant transformacions posteriors.

S'utilitza l'opció APPEND, amb la que s'indica que els registres s'insereixin a continuació de les dades que ja hi ha a la taula. S'indica també que, si hi ha camps sense valors, es posin a NULL (això ho indica l'opció TRAILING NULLCOLS).

Una transformació prèvia necessària és la de reemplaçar els punts dels decimals per comes. Sqlloader no entén que un decimal amb punt és un número. D'aquesta manera no cal fer la transformació de format de dades. A més, quan a un camp no hi ha dades, s'omple el mateix amb dos punts. Sqlloader dóna l'opció d'indicar que si el camp a guardar a la base de dades conté dos punts ho deixi a NULL. Exemple: *Asal2003 DECIMAL EXTERNAL NULLIF Asal2003='..'.*

Amb l'execució de sqlloader també queda controlada l'eliminació dels "totals" aportats per algunes CCAA, ja que permet triar quines columnes del fitxer són les que es volen guardar a la base de dades.

Per automatitzar més la càrrega de dades, s'ha creat el fitxer per lots *CarregaCTLs.bat*, que executa la carrega de tots els fitxers.

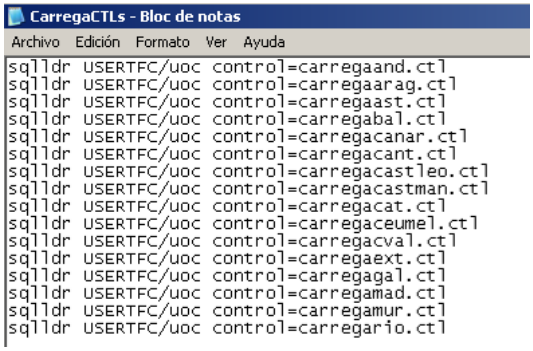


Figura 11.- Fitxer per lots per executar la càrrega de dades tributàries

Amb aquesta càrrega de dades s'aconsegueix tenir a la taula temporal TRIBUTACIO el contingut de tots els fitxers aportats per l'OADT. No és necessari realitzar cap procés especial per extreure les dades de CEUTA i MELILLA, ja que amb aquest algorisme de càrrega totes les dades es troben de manera seqüencial a la taula TRIBUTACIO.

El procés de càrrega i de transformació de dades s'automatitza al màxim possible, executant un algorisme en PL/SQL que fa la càrrega a les taules definitives i realitza els processos de transformació de dades.

Aquest algorisme fa servir dues taules temporals:

- TRIBUTACIO: guarda tots els registres carregats amb sqlloader de totes les CCAA, un a continuació de l'altre. Així es pot tractar aquesta taula de manera seqüencial i amb una passada es tenen les dades, amb les transformacions adients, a la taula definitiva PRESTACIO.

- TEMPREST: s'utilitza per guardar cada paquet de dades de què es composen les instàncies del fet prestacions, per un any i un tipus de prestació en concret. Aquesta taula s'omple amb la iteració del procés en PL/SQL que realitza la càrrega i, quan es guarden les prestacions a la taula definitiva PRESTACIO, es buida per poder continuar carregant les següents dades.

El funcionament de l'algorisme bàsicament consisteix a recórrer, mitjançant un cursor, tots els registres de TRIBUTACIO, identificant primerament a quina CA pertanyen les dades (els noms de les CA es guarden en majúscules per unificar els formats). Després es busquen totes les dades referents a SALARIOS d'aquesta CA que són d'interès (persones, retribucions, etc), i s'ignoren aquelles dades de "totals" aportades per algunes CCAA. Quan es tenen totes les dades es realitzen els següents processos:

- Si de retribucions i/o retencions només hi ha valors mitjans, es fan càlculs per obtenir els valors absoluts.

- Si no es té informació tributària dels anys 2002/2003, s'omplen les dades amb la mitjana de la resta d'anys de què es disposa.

- Si algun valor és NULL, es posa a 0.

- S'insereixen les dades a la taula PRESTACIO i s'esborra la taula TEMPPREST per continuar l'algorisme.

A continuació, es fa el mateix procés per PENSIONES i DESEMPLEOS. Quan s'acaba amb una CA es continua amb la següent.

Quan s'acaba la carrega a la taula definitiva es tenen un total de 1632 registres.

A continuació, es pot veure el pseudocodi de l'algorisme de càrrega i transformació de dades tributàries per omplir la taula PRESTACIO.

```
//mitjançant un cursor es llegeixen els registres de la taula TRIBUTACIO (que és la taula temporal on estan les dades de tots els fitxers CSV amb informació tributària)
Llegir registre
//bucle per recórrer tota la taula TRIBUTACIO
Mentre quedin registres repetir
        //TEMPPREST guarda informació completa d'una instància del FET principal
        Esborrar taula TEMPPREST
        Guardar nom CA
        //bucle per arribar a SALARIOS
        Mentre no s'arribi a informació de SALARIOS o final de la taula repetir
                Llegir registre
        Fi mentre
        //bucle per guardar informació relativa a SALARIOS
        Mentre no s'arribi a informació de PENSIONES o final de la taula repetir
                Llegir registre
                Si registre és de PERSONAS
                        Guardar a TEMPPREST totes les dades del registre actual de PERSONAS
                Sino  Si registre és de RETRIBUCIONES
                        //es sap que existeix informació de les retribucions, per després controlar si calen fer càlculs per obtenir dades a partir dels valors mitjans
                        ExisteixRetribucio = TRUE
                        Guardar a TEMPPREST totes les dades del registre actual de RETRIBUCIONES
                    Sino Si registre és de RETRIBUCION MEDIA ANUAL i ExisteixRetribucio == FALSE
                                //no hi havia dades de retribució absolutes, cal guardar els valors mitjans aportats
                                Guardar a TEMPPREST totes les dades del registre actual de RETRIBUCION MEDIA ANUAL
                                Sino Si registre és de RETENCIONES
                                        //es sap que existeix informació de les retencions, per després controlar si calen fer càlculs per obtenir dades a partir dels valors mitjans
                                        ExisteixRetencio = TRUE
                                        Guardar a TEMPPREST totes les dades del registre actual de RETENCIONES
                                    Sino Si registre és de TIPO MEDIO DE RETENCION i ExisteixRetencio == FALSE
                                                //no hi havia dades de retenció absolutes, cal guardar els valors mitjans aportats
                                                Guardar a TEMPPREST totes les dades del registre actual de TIPO MEDIO DE RETENCION
                                            Fi si
                                        Fi si
                                Fi si
                        Fi si
                Fi si
        Fi repetir
  //Control de línies de tributació per aquelles CCAA que només aporten valor mitjans i no absoluts
  //si no es té el valor de retribució absolut
Si ExisteixRetribucio == false
        //es guarda a la taula temporal el càlcul del valor absolut a partir dels valors mitjans aportats per la CA
        Guardar a array_temporal1 el registre de línea PERSONAS
        Guardar a array_temporal2 el registre de línea RETRIBUCION MEDIA ANUAL
```

```
        Inserir a TEMPPREST el producte dels array_temporal1 i array_temporal2
Fi si
//si no es té el valor de retenció absolut
Si ExisteixRetencio == false
        //es guarda a la taula temporal el càlcul del valor absolut a partir dels valors mitjans aportats per la CA
        Guardar a array_temporal1 el registre de línea RETRIBUCIONES
        Guardar a array_temporal2 el registre de línea TIPO MEDIO DE RETENCION
        Inserir a TEMPPREST el producte dels array_temporal1 i array_temporal2/100
Fi si
//ja es tenen les dades absolutes de RETRIBUCIONES i RETENCIONES, s'esborren les mitjanes
Esborrar de TEMPPREST línea amb dades de TIPO MEDIO DE RETENCION
Esborrar de TEMPPREST línea amb dades de RETRIBUCION MEDIA ANUAL
//es posen a FALSE les variables de control
ExisteixRetribucio = FALSE
ExisteixRetencio = FALSE
//control d'aquelles comunitats que no aporten informació dels anys 2002/2003, es decideix omplir les dades NULL amb
les mitjanes de la resta d'anys
Guardar a variables temporals registre de persones de TEMPPREST
//si no hi ha dades del 2002 de PERSONAS
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de PERSONAS
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre PERSONAS
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de PERSONAS
                Calcular mitjana de 2003 a 2009 de tots els camps del registre PERSONAS
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//si no hi ha dades del 2002 de RETRIBUCIONES
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de RETRIBUCIONES
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre RETRIBUCIONES
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de RETRIBUCIONES
                Calcular mitjana de 2003 a 2009 de tots els camps del registre RETRIBUCIONES
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//si no hi ha dades del 2002 de RETENCIONES
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de RETENCIONES
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre RETENCIONES
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de RETENCIONES
                Calcular mitjana de 2003 a 2009 de tots els camps del registre RETENCIONES
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//ja es troben les dades de forma adequada, sense mitjanes i amb dades completes del 2002 al 2009
//s'omplen les dades definitives de SALARIOS a la taula de PRESTACIO
Selecció de dades dels camps dels anys 2002 a 2009 de PERSONAS, RETRIBUCIONES, RETENCIONES de la taula TEMPPREST
```

```
Inserir a taula PRESTACIO dades anteriors amb CA, tipus de prestació i tipus de retribució adients
////S'arriba a les dades de PENSIONES de la CA que s'estan introduint a la base de dades
 //s'esborra la informació de SALARIS que ja s'ha guardat a la taula definitiva
……..
/////El codi anterior s'executa de nou per obtenir les dades referents a PENSIONES, no es posa per reduir el pseudocodi
……..
////S'arriba a les dades de DESEMPLEO de la CA que s'estan introduint a la base de dades
 //s'esborra la informació de PENSIONES que ja s'ha guardat a la taula definitiva
Esborrar dades de TEMPPREST
        //bucle per guardar informació relativa a DESEMPLEO, repetir mentre no es trobi el nom de la següent CA a tractar o s'acabin els registres de la taula TRIBUTACIO
        //variable per controlar si el final de les dades és perquè encara queden comunitats a tractar
        esComunitat = FALSE
        Mentre esComunitat = FALSE o final de la taula repetir
                q = 1
                Llegir registre
                //bucle per comprovar si el que s'està tractant al registre actual és el nom de la nova CA o és un registre de dades d'una prestació. Es compara el registre amb tots els noms de CCAA
                Mentre esComunitat = TRUE o q != 18
                        Si registre == array_comunitats (q)
                                esComunitat = TRUE //s'ha trobat una CA nova
                        Fi si
                        q = q+1
                Fi mentre
                Si registre és de PERSONAS
                        Guardar a TEMPPREST totes les dades del registre actual de PERSONAS
                Sino Si registre és de RETRIBUCIONES
                        //es sap que existeix informació de les retribucions, per després controlar si calen fer càlculs per obtenir dades a partir dels valors mitjans
                        ExisteixRetribucio = TRUE
                        Guardar a TEMPPREST totes les dades del registre actual de RETRIBUCIONES
                        Sino Si registre és de RETRIBUCION MEDIA ANUAL i ExisteixRetribucio == FALSE
                                //no hi havia dades de retribució absolutes, cal guardar els valors mitjans aportats
                                Guardar a TEMPPREST totes les dades del registre actual de RETRIBUCION MEDIA ANUAL
                                Sino Si registre és de RETENCIONES
                                        //es sap que existeix informació de les retencions, per després controlar si calen fer càlculs per obtenir dades a partir dels valors mitjans
                                        ExisteixRetencio = TRUE
                                        Guardar a TEMPPREST totes les dades del registre actual de RETENCIONES
                                    Sino Si registre és de TIPO MEDIO DE RETENCION i ExisteixRetencio == FALSE
                                                //no hi havia dades de retenció absolutes, cal guardar els valors mitjans aportats
                                                Guardar a TEMPPREST totes les dades del registre actual de TIPO MEDIO DE RETENCION
                                            Fi si
                                    Fi si
                                Fi si
                            Fi si
                        Fi si
        Fi repetir
//Control de línies de tributació per aquelles CCAA que només aporten valor mitjans i no absoluts
//si no es té el valor de retribució absolut
Si ExisteixRetribucio == false
        //es guarda a la taula temporal el càlcul del valor absolut a partir dels valors mitjans aportats per la CA
        Guardar a array_temporal1 el registre de línea PERSONAS
        Guardar a array_temporal2 el registre de línea RETRIBUCION MEDIA ANUAL
```

```
        Inserir a TEMPPREST el producte dels array_temporal1 i array_temporal2
Fi si
//si no es té el valor de retenció absolut
Si ExisteixRetencio == false
        //es guarda a la taula temporal el càlcul del valor absolut a partir dels valors mitjans aportats per la CA
        Guardar a array_temporal1 el registre de línea RETRIBUCIONES
        Guardar a array_temporal2 el registre de línea TIPO MEDIO DE RETENCION
        Inserir a TEMPPREST el producte dels array_temporal1 i array_temporal2/100
Fi si
//ja es tenen les dades absolutes de RETRIBUCIONES i RETENCIONES, s'esborren les mitjanes
Esborrar de TEMPPREST línea amb dades de TIPO MEDIO DE RETENCION
Esborrar de TEMPPREST línea amb dades de RETRIBUCION MEDIA ANUAL
//es posen a FALSE les variables de control
ExisteixRetribucio = FALSE
ExisteixRetencio = FALSE
//control d'aquelles CCAA que no aporten informació dels anys 2002/2003, es decideix omplir les dades NULL amb les mitjanes de la resta d'anys
Guardar a variables temporals registre de persones de TEMPPREST
//si no hi ha dades del 2002 de PERSONAS
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de PERSONAS
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre PERSONAS
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de PERSONAS
                Calcular mitjana de 2003 a 2009 de tots els camps del registre PERSONAS
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//si no hi ha dades del 2002 de RETRIBUCIONES
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de RETRIBUCIONES
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre RETRIBUCIONES
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de RETRIBUCIONES
                Calcular mitjana de 2003 a 2009 de tots els camps del registre RETRIBUCIONES
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//si no hi ha dades del 2002 de RETENCIONES
Si variable temporal de 2002 = NULL
        //si no hi ha dades del 2003 de RETENCIONES
        Si variable temporal de 2003 = NULL
                Calcular mitjana de 2004 a 2009 de tots els camps del registre RETENCIONES
                Actualitzar registres de 2002 i 2003  a TEMPPREST
        Si no
        //hi ha dades del 2003 però no del 2002 de RETENCIONES
                Calcular mitjana de 2003 a 2009 de tots els camps del registre RETENCIONES
                Actualitzar registres de 2002  a TEMPPREST
        Fi si
Fi si
//ja es troben les dades de forma adequada, sense mitjanes i amb dades completes del 2002 al 2009
//s'omplen les dades definitives de DESEMPLEO a la taula de PRESTACIO
```

Selecció de dades dels camps dels anys 2002 a 2009 de PERSONAS, RETRIBUCIONES, RETENCIONES de la taula TEMPPREST

Inserir a taula PRESTACIO dades anteriors amb CA, tipus de prestació i tipus de retribució adients

Fi mentre

Tancar cursor

Fi

Es pot consultar el script C*arregaPrestacio.sql* per veure el codi PL/SQL.

Un cop es tenen les dades tributàries carregades a la taula definitiva PRESTACIO, es procedeix a realitzar la càrrega de les dades de població i població activa. El procés de càrrega de dades dels fitxers CSV és similar a l'anterior. S'executa el fitxer per lots *CarregaPobCA.bat* i es carreguen les dades de manera automàtica. S'utilitzen les següents taules temporals:

- TMPCOMAUTON: guarda el contingut de tots els fitxers amb les dades de població de tots els anys, un darrere de l'altre. Es pot saber on acaben les dades d'un any i comencen les del següent perquè sempre hi ha 17 registres (un per cada CA).
- TMPCOMAUTONPERC: guarda el contingut del fitxer amb les dades de població activa.

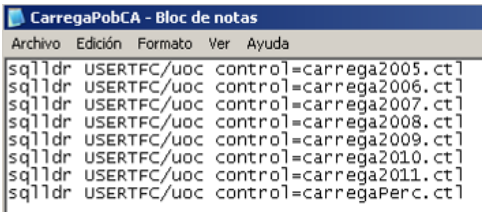


Figura 12.- Fitxer per lots per executar la càrrega de població i població activa

Un cop es tenen les dades carregades a les taules temporals, el procés de transformació i de càrrega a la taula POBLACIO, que serà la definitiva, també s'automatitza amb un altre algorisme en PL/SQL.

El funcionament de l'algorisme consisteix a recórrer, mitjançant un cursor, tots els registres de TEMPCOMAUTON. S'agafen les dades de població activa de TEMPCOMAUTONPERC, per cada any i CA, i es guarden al mateix registre de la taula definitiva.

En el mateix algorisme es realitzen les modificacions de les dades requerides als processos ETL:

- S'eliminen les dades de Navarra i País Basc de les taules TMPCOMAUTON i TMPCOMAUTONPERC, ja que tenen un règim foral propi.
- S'unifiquen els noms de les CCAA, ja que no tots els fitxers font contenen el mateix nom descriptiu per una mateixa CA.
- Es controla que si un camp és NULL, es guardi un 0.

Amb les dades aportades hi han 119 registres a la taula de dades definitiva, dades de 7 anys per 17 CCAA.

A continuació es pot veure el pseudocodi de l'algorisme de carrega i transformació de les dades de població a la taula POBLACIO.

```
//s'esborra de la taula de dades de CCAA i de la de població activa temporals les dades de Navarra i el País Basc, ja que no hi ha dades tributàries d'aquestes
Esborrar de TMPCOMAUTON quan comaut = 'Navarra (Comunidad Foral de)' o comaut ='País Vasco'
//s'actualitzen els noms de les CCAA, per unificar com estan escrits amb els fitxers d'informació tributària.
Actualitzar TMPCOMAUTON i TMPCOMAUTONPERC posar comaut = 'CEUTA' quan comaut = 'Ciudad autónoma de Ceuta';
.......
//es fa el mateix procés d'actualització amb totes les CCAA
......
Actualitzar TMPCOMAUTON posar comaut = 'ARAGÓN' quan comaut = 'Aragón';
//mitjançant un cursor es llegeixen els registres de la taula TMPCOMAUTON (que és la taula temporal on estan les dades de tots els fitxers CSV amb informació de població)
//bucle principal per recórrer les dades de població per CA dels anys 2005-2011
Mentre quedin registres per llegir i no s'arribi a l'any 2012
        //bucle per recórrer tota la taula amb la informació de les dades de població de tots els anys
        Mentre quedin registres per llegir i encara no s'hagin recorregut totes les CCAA d'aquest any
                Llegir registre
                Guardar id de CA segons el nom trobat al registre
                Guardar id de l'any del que s'està tractant les dades
//es guarda la informació del percentatge de població activa del mateix any a una variable (dades extretes de la taula temporal amb la població activa TMPCOMAUTONPERC)
                En cas que
                        Any sigui 2005, guardar a variablePercentatge dades de població activa del 2005
                        Any sigui 2006, guardar a variablePercentatge dades de població activa del 2006
                        Any sigui 2007, guardar a variablePercentatge dades de població activa del 2007
                        Any sigui 2008, guardar a variablePercentatge dades de població activa del 2008
                        Any sigui 2009, guardar a variablePercentatge dades de població activa del 2009
                        Any sigui 2010, guardar a variablePercentatge dades de població activa del 2010
                        Any sigui 2011, guardar a variablePercentatge NULL (hi ha dades d'aquest any)
                Fi cas
//es guarda a la taula definitiva de dades de població la informació de població activa i habitants. Taula POBLACIO
                Inserir en POBLACIO dades extretes a la iteració del bucle
 Fi mentre
Fi Mentre
//tancar cursor
Fi
```

Es pot consultar el script *CarregaCursorsPobCA.sql* per veure el codi PL/SQL.

Per últim, es crea un algorisme de modificació del nombre d'habitants de cada CA perquè les dades siguin més realistes. Es calcula com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent. L'algorisme el que fa és llegir cada registre de la taula POBLACIO i fa la mitjana d'homes i dones amb l'any següent, actualitzant així tots els valors.

A continuació es pot veure el pseudocodi de l'algorisme de modificació.

```
//bucle principal per controlar que es modifiquin tots els registres de població dels anys 2005-2010 i de totes les CCAA
Mentre no s'hagin recorregut tots els anys
        //bucle per modificar els registres de totes les CCAA d'un mateix any
        Mentre no s'hagin recorregut totes les CA per aquest any
                //es guarden les dades de població de l'any a variables
```

```
                    Seleccionar nombre d'homes i dones de la taula POBLACIO per la CA i Any que s'estan tractant a
                    aquesta iteració
                    //es guarden les dades de població de l'any següent a variables
                    Seleccionar nombre d'homes i dones de la taula POBLACIO per la CA que s'està tractant en aquesta
                    iteració i l'any posterior
                    //s'actualitzen les dades de població amb la mitjana dels dos anys
                    Actualitzar registre d'homes de la taula POBLACIO
                    Actualitzar registre de dones de la taula POBLACIO
                Fi mentre
            Fi mentre
            Fi
```

Es pot consultar el script *ModificaValorsPobCA.sql* per veure el codi PL/SQL.

Un cop realitzat el procés de càrrega ja es poden eliminar les taules temporals.

3.6. Creació de les consultes

A continuació s'indica el procés de creació de les consultes, amb l'explicació detallada de cada una d'elles.

3.6.1. Creació del model amb Discoverer

Com s'indica a les especificacions del projecte, les consultes i informes finals es realitzen mitjançant Oracle Discover. Aquesta eina facilita a l'usuari la creació de noves consultes i informes, a partir del model dissenyat, d'una manera àgil i senzilla.

La part de creació de l'àrea de negoci es fa amb Oracle Discoverer Administrator. Les consultes, amb el corresponent llibre de treball, amb Oracle Discoverer Desktop. L'accés a aquestes dues aplicacions es fa amb els mateixos usuari i paraula de pas que per la creació de les taules de la base de dades en Oracle (Usuari: **USERTFC** i Paraula de pas: **uoc**).

3.6.1.1. Creació de l'àrea de negoci

Per poder treballar amb un àrea de negoci cal crear un EUL (de l'anglès, End User Layer). Aquest EUL, que és la capa d'usuari final, és per aquell que realitza el procés de creació. Posteriorment es crea l'àrea de negoci, anomenada AreaNegoci_OADT, amb les taules i dades necessàries per poder efectuar els informes. A més, es defineixen les relacions (unions) necessàries entre els elements que s'utilitzen. A l'apartat de les consultes s'expliquen altres carpetes que s'incorporen a l'àrea de negoci per poder realitzar determinats informes.

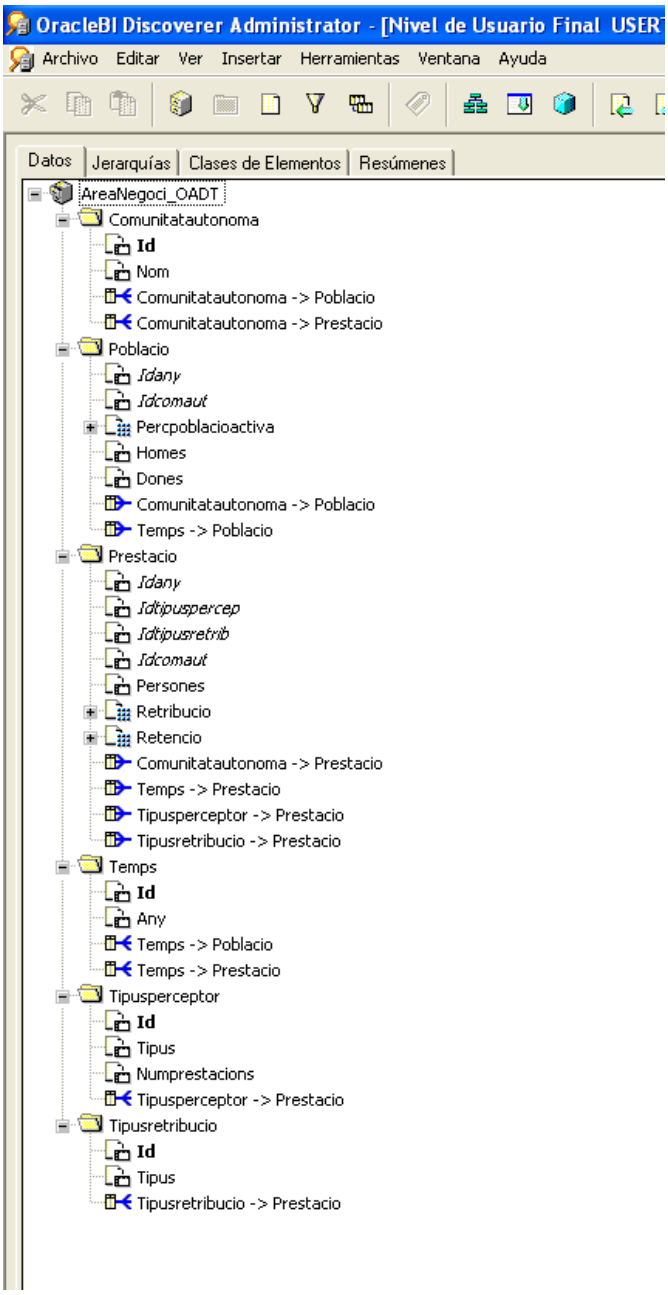


Figura 13.- Àrea de negoci

3.6.1.2. Creació del llibre de treball

Un cop creat l'àrea de negoci, es poden crear les consultes amb Oracle Discoverer Desktop.

S'executa Oracle Discoverer Desktop amb l'usuari propietari de l'EUL creat anteriorment. Sobre l'AreaNegoci_OADT es crea un nou llibre de treball, que és com el Discoverer organitza les consultes. Cada llibre pot contenir múltiples fulls, cadascun d'ells correspon a una consulta.

El llibre creat, amb tots els informes, s'anomena LlibreOADT i es troba a UOCPRACT.vdi, path E:\fitxers.

3.6.2. Consultes creades

A continuació es poden veure les 12 consultes creades. Hi ha una que no es requereix pel client però es creu útil.

Cada full del llibre de treball conté una consulta, excepte per la projecció d'indicadors que s'utilitzen dues. A més, en ells es veuen les condicions, càlculs, valors totals i formats utilitzats a cada consulta.

3.6.2.1. Informe total retribució

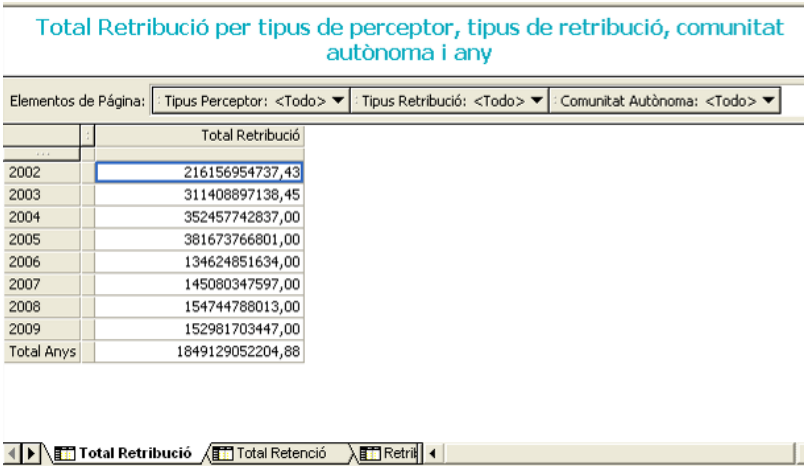


Figura 14.- Informe total retribució

Informe on es veu l'import de les retribucions de cada prestació. Es pot filtrar per tipus de perceptor, tipus de retribució i CA. El resultat es veu per tots els anys a estudi. S'incorpora el total de sumar les retribucions de tots els anys.

3.6.2.2. Informe total retenció

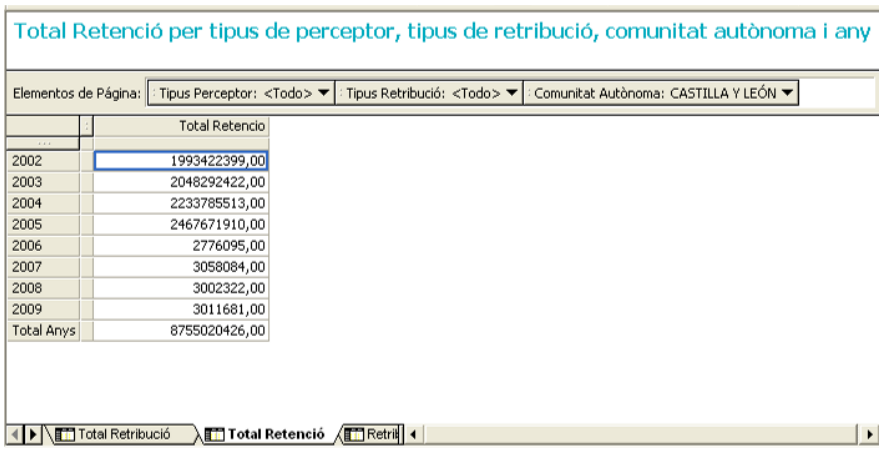


Figura 15.- Informe total retenció

Informe on es veu l'import de les retencions de cada prestació. Es pot filtrar per tipus de perceptor, tipus de retribució i CA. El resultat es veu per tots els anys a estudi. S'incorpora el total de sumar les retencions de tots els anys.

3.6.2.3. Informe retribució mitjana

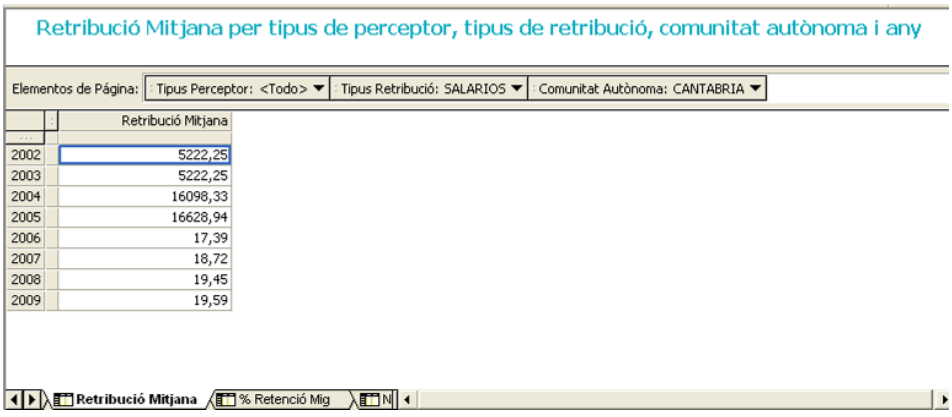


Figura 16.- Informe retribució mitjana

Informe on es veu la retribució mitjana. Aquest càlcul es realitza dividint l'import de la retribució entre les persones que ho reben. És necessari posar al dividend el sumatori de la retribució, per realitzar correctament els càlculs de les retribucions mitjanes de diverses prestacions. Es té en compte que, si el perceptor rep més d'una prestació, es considera com si fossin més d'una persona pel càlcul total.

A l'informe es pot filtrar per tipus de perceptor, tipus de retribució i CA. El resultat es veu per tots els anys a estudi.

3.6.2.4. Informe percentatge de retenció mig

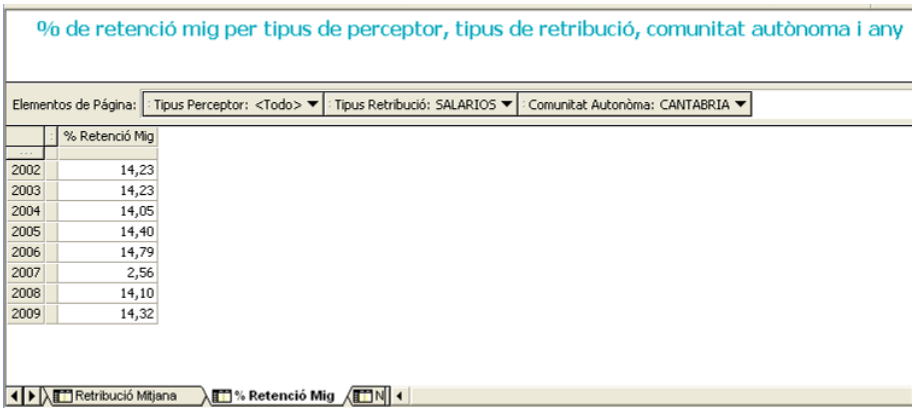


Figura 17.- Informe percentatge de retenció mig

Informe on es veu el percentatge de retenció mig. Aquest càlcul es realitza dividint l'import retingut entre la retribució i fent el tant per cent. És necessari posar al dividend la mitjana de les retencions i al divisor la mitjana de les retribucions, per poder realitzar correctament els càlculs quan hi ha diversos operands i es vol el percentatge de retenció entre elles.

A l'informe es pot filtrar per tipus de perceptor, tipus de retribució i CA. El resultat es veu per tots els anys a estudi.

3.6.2.5. Número de retribucions mig

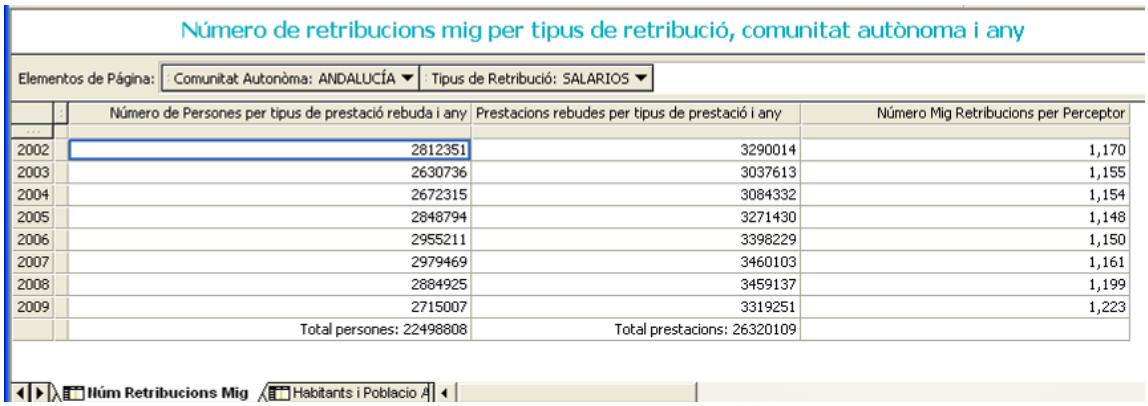


Figura 18.- Informe número de retribucions mig

En aquesta consulta cal indicar que a l'enunciat especifica que si una persona rep pensió i salari (per exemple) compta com dues. S'entén que vol dir que una persona rep dues prestacions.

Per poder realitzar el càlcul de retribucions mitjanes, i que sigui real quan es calculen els totals, es considera necessari dividir el número de persones que reben una prestació entre el número de prestacions que rep aquest tipus de perceptor, ja que sinó comptarien doble o triple i el càlcul sempre donaria una relació 1/1. S'utilitza l'atribut NUMPRESTACIONS de la taula TIPUSPERCEPTOR per fer els càlculs.

A més, és necessari realitzar el sumatori de la divisió per efectuar els càlculs totals. El número de Retribucions és el total de persones que estan dins d'un segment segons el tipus de perceptor que sigui (la columna de l'informe *Prestacions rebudes per tipus de prestació i any*). S'inclouen els sumatoris dels totals.

La columna *Número Mig Retribucions per Perceptor* es calcula com la divisió de la columna *Prestacions Rebudes per tipus de prestació i any* entre *Número de Persones per Tipus de Prestació Rebuda i Any.*

3.6.2.6. Percentatge de població per segment

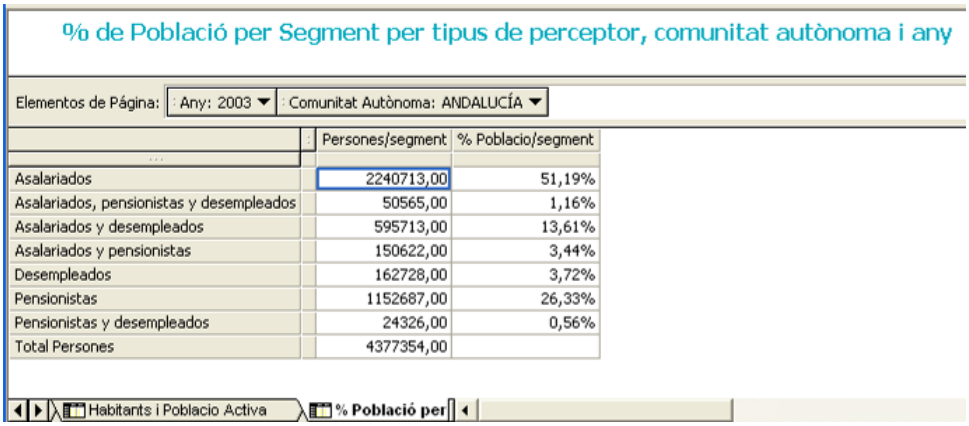


Figura 19.- Informe percentatge de població per segment

En aquest informe es realitza el càlcul del percentatge de població per cada tipus de perceptor. La columna Persones/segment indica quantes persones hi ha de cada tipus de perceptor. El càlcul es realitza dividint el número total de persones entre el número de prestacions que hi reben. Això és degut al fet que hi ha persones que cobren més d'una retribució i, si no es fa la divisió, es tindrien en compte més d'una vegada.

La columna de % Poblacio/segment és el percentatge de la columna anterior.

Es poden consultar les dades per any, CA i tipus de perceptor.

3.6.2.7. Nombre de treballadors/Nombre de perceptors no actius (determina la sostenibilitat del sistema)

En aquesta consulta es calcula la proporció entre el nombre de treballadors i el nombre de perceptors no actius. Es considera que els perceptors no actius són els aturats i jubilats, ja que són perceptors de retribucions i influeixen a la sostenibilitat del sistema. S'entén que és diferent aquest concepte del de població no activa, que només inclou els jubilats.

Per realitzar aquest informe cal filtrar la taula de prestacions pel tipus de retribució. Es necessiten dos filtres diferents pels mateixos camps, ja que el nombre de treballadors (dividend de l'operació que s'ha de realitzar) necessita que el tipus de retribució sigui SALARIOS i perceptors no actius necessita que sigui PENSIONES i DESEMPLEO.

Es crea una vista materialitzada, anomenada MVPERCEPTORS, per fer més senzilles les tasques de filtratge, de manera que es tenen les dues columnes a la vista amb cada filtre aplicat. A més, s'incorporen més camps per futures consultes.

El codi sql de la consulta que utilitza la vista és el següent:

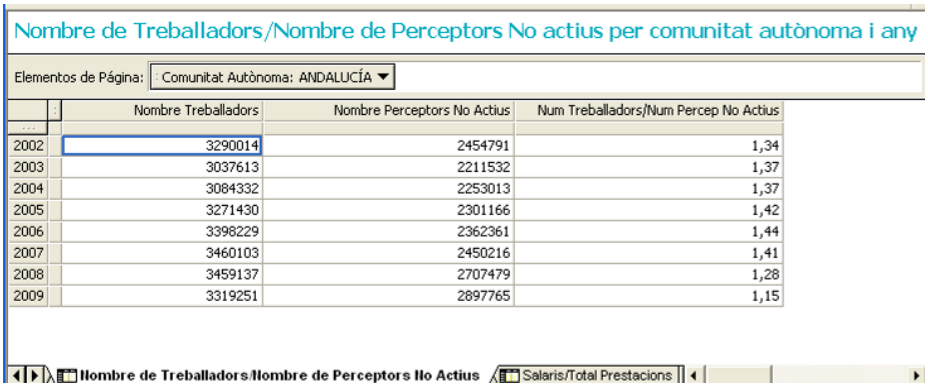
```
AS WITH
T1 AS (
        select sum(p1.persones)NOACTIUS, sum(p1.retribucio)RETRIBUCIONSNOACTIUS, p1.idany, p1.idcomaut
        from prestacio p1
        where (p1.idtipusretrib = 2 or p1.idtipusretrib = 3)
        group by  p1.idany, p1.idcomaut
        order by 3,4 ),

T2 AS (
        select sum(p2.persones)ACTIUS, sum(p2.retribucio)RETRIBUCIONSACTIUS, p2.idany, p2.idcomaut
        from prestacio p2
        where (p2.idtipusretrib = 1)
        group by  p2.idany, p2.idcomaut
        order by 3,4)

SELECT ACTIUS, NOACTIUS, RETRIBUCIONSACTIUS, RETRIBUCIONSNOACTIUS, T1.IDANY, T1.IDCOMAUT FROM T1,T2
WHERE T1.IDANY = T2.IDANY AND T1.IDCOMAUT = T2.IDCOMAUT ;
```

Amb aquesta vista es tenen la relació de persones actives (assalariats), no actives (pensionistes i aturats) i les retribucions que reben cadascun d'ells.

Column Name	Data Type
ACTIUS	NUMBER
NOACTIUS	NUMBER
RETRIBUCIONSACTIUS	NUMBER
RETRIBUCIONSNOACTIUS	NUMBER
IDANY	NUMBER(5,0)
IDCOMAUT	NUMBER(5,0)

Figura 20.- vista materialitzada MVPERCEPTORS

S'incorpora la vista a una nova carpeta de l'àrea de negoci i es creen unions, per relacionar la vista amb les taules que es necessiten per realitzar els informes.


Nombre de Treballadors/Nombre de Perceptors No actius per comunitat autònoma i any

Elementos de Página: Comunitat Autònoma: ANDALUCÍA

	Nombre Treballadors	Nombre Perceptors No Actius	Num Treballadors/Num Percep No Actius
2002	3290014	2454791	1,34
2003	3037613	2211532	1,37
2004	3084332	2253013	1,37
2005	3271430	2301166	1,42
2006	3398229	2362361	1,44
2007	3460103	2450216	1,41
2008	3459137	2707479	1,28
2009	3319251	2897765	1,15

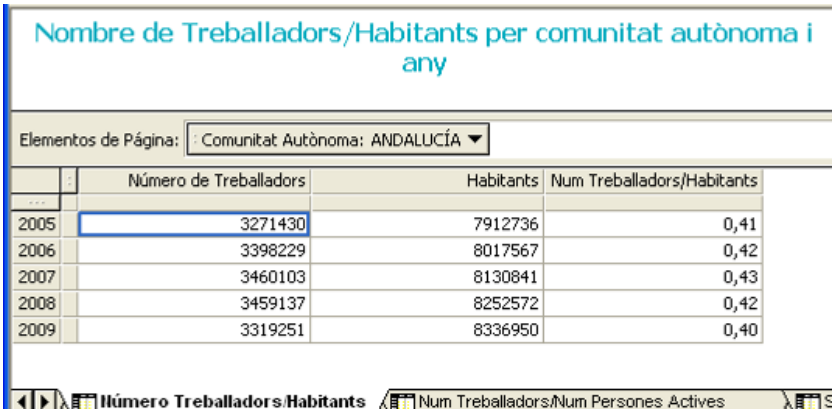
Nombre de Treballadors/Nombre de Perceptors No Actius | Salaris/Total Prestacions

Figura 21.- Informe Nombre de treballadors/Nombre de Perceptors No actius

El filtratge per seleccionar Nombre de treballadors i Nombre Perceptors No Actius es realitza mitjançant la vista. No cal cap filtre addicional a Discoverer.

Per obtenir la relació d'aquests paràmetres es fa la divisió de la columna *Nombre Treballadors* entre *Nombre Perceptors No Actius*.

3.6.2.8. Nombre de treballadors/Habitants


Nombre de Treballadors/Habitants per comunitat autònoma i any

Elementos de Página: Comunitat Autònoma: ANDALUCÍA

	Número de Treballadors	Habitants	Num Treballadors/Habitants
2005	3271430	7912736	0,41
2006	3398229	8017567	0,42
2007	3460103	8130841	0,43
2008	3459137	8252572	0,42
2009	3319251	8336950	0,40

Número Treballadors/Habitants | Num Treballadors/Num Persones Actives

Figura 22.- Informe Nombre de treballadors/Habitants

En aquest informe es realitza el càlcul de la proporció entre el nombre de persones que treballen (reben SALARIS) i el nombre d'habitants. S'observa que els anys d'estudi són del 2005 al 2009, ja que són dels anys que es tenen dades de població.

El número de treballadors es pot obtenir de la taula de prestacions o de la vista materialitzada MVPERCEPTORS. S'ha realitzat l'informe amb la segona opció. El nombre d'habitants s'obté de la taula POBLACIO (el fet secundari del disseny del model de dades), sumant homes i dones.

3.6.2.9. Nombre de treballadors/Número de persones actives

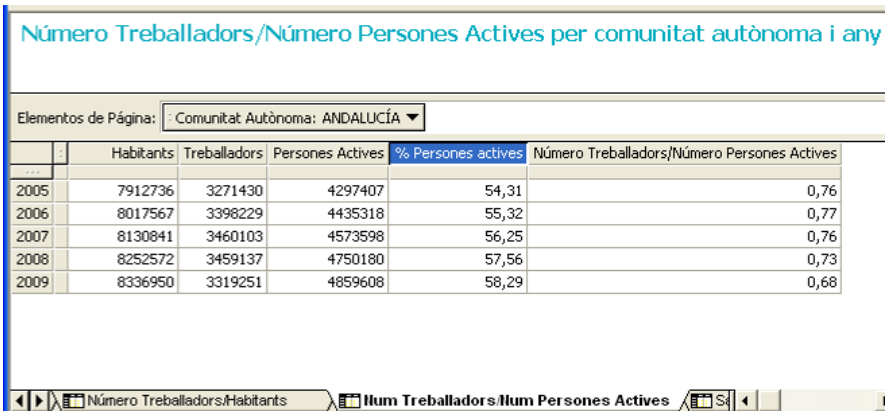


Figura 23.- Informe Nombre de treballadors/Número de persones actives

En aquest informe es realitza el càlcul de la proporció entre el nombre de persones que treballen (reben SALARIS) i el nombre de persones actives (que treballen o estan a l'atur). S'observa que els anys d'estudi són del 2005 al 2009, ja que són dels anys que es tenen dades de població.

El número d'habitants es calcula com la suma dels valors homes i dones guardats a la taula POBLACIO. El número de persones actives surt a partir del percentatge de població activa aportat per les CCAA i guardat també a la taula POBLACIO. El número de treballadors surt de la vista materialitzada MVPERCEPTORS.

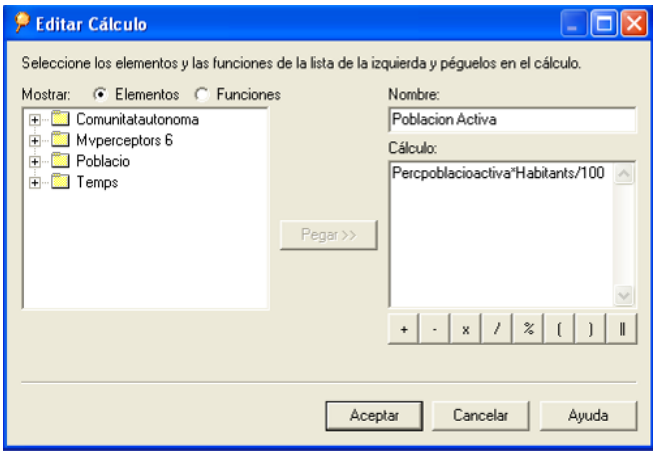
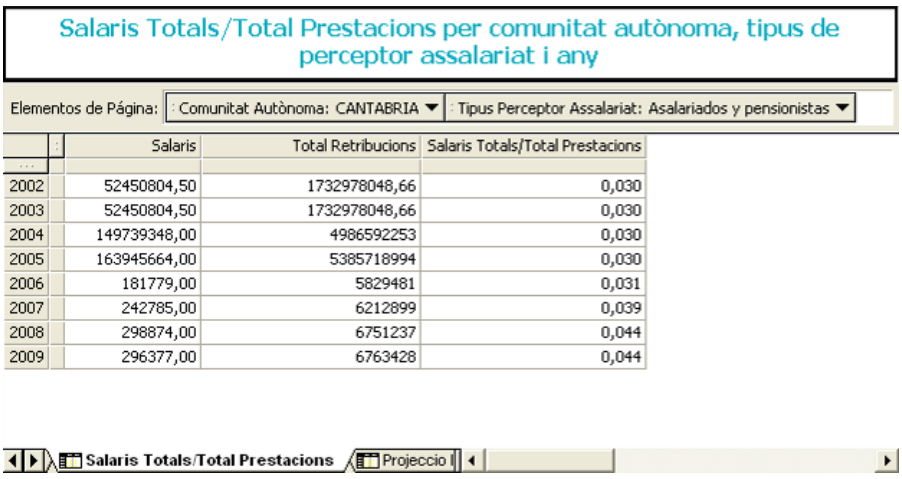


Figura 24.- càlcul població activa

3.6.2.10. Salaris Totals/Total Prestacions



Salaris Totals/Total Prestacions per comunitat autònoma, tipus de perceptor assalariat i any

Elementos de Página: Comunitat Autònoma: CANTABRIA | Tipus Perceptor Assalariat: Asalariados y pensionistas

	Salaris	Total Retribucions	Salaris Totals/Total Prestacions
2002	52450804,50	1732978048,66	0,030
2003	52450804,50	1732978048,66	0,030
2004	149739348,00	4986592253	0,030
2005	163945664,00	5385718994	0,030
2006	181779,00	5829481	0,031
2007	242785,00	6212899	0,039
2008	298874,00	6751237	0,044
2009	296377,00	6763428	0,044

Salaris Totals/Total Prestacions | Projeccio

Figura 25.- Informe Salaris Totals/Total Prestacions

En aquest informe es realitza el càlcul de la proporció entre el total de salaris i el nombre total de prestacions. Es pot consultar de forma agregada per CA, tipus de perceptor i any. S'imposa la condició que el tipus de retribució sigui sempre SALARI.

A continuació es veu el càlcul per obtenir el total de retribucions i la condició imposada.

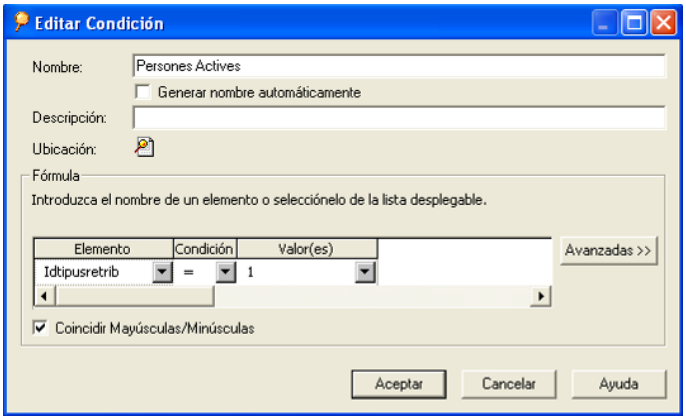


Figura 26.- condició de l'informe

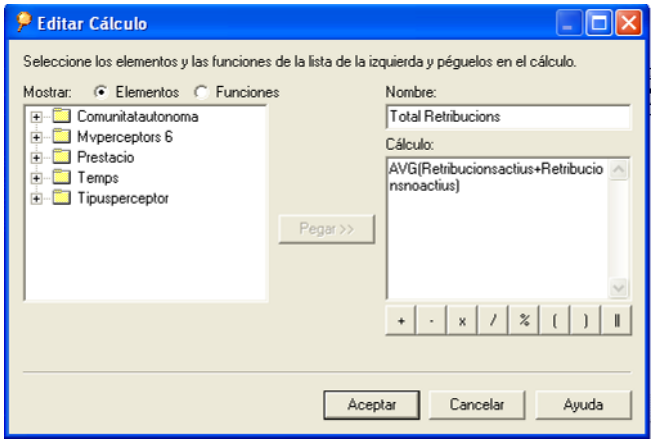


Figura 27.- càlcul del total de retribucions

3.6.2.11. Habitants i població activa

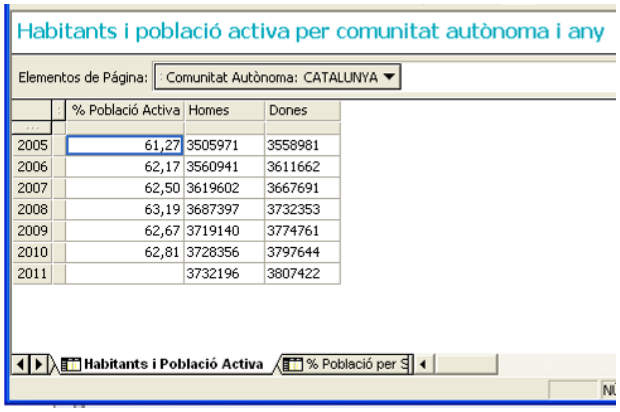


Figura 28.- Informe Habitants i població activa

Aquest informe és addicional. Mostra informació de la taula POBLACIO. Es tenen els habitants de cada CA, separat per sexes i el percentatge de població activa.

3.6.2.12. Projecció dels indicadors *NumTreb/NumPerceptors no actius* i *NumTreb/Poblacio* per determinar el punt de col·lapse

Gràfica dels indicadors

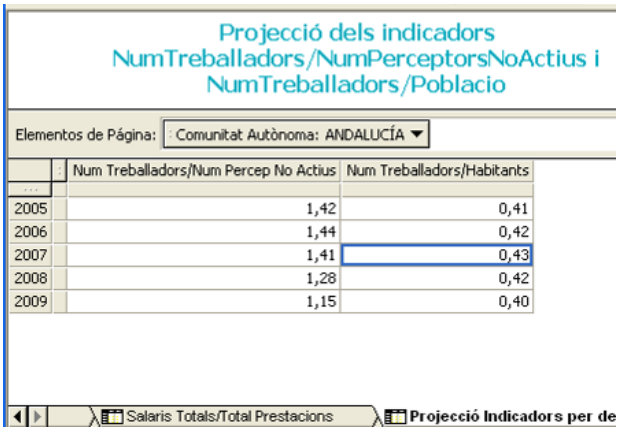
S'entén per població perceptora no activa a aturats i jubilats, ja que reben prestacions.

Aquest informe es realitza utilitzant dos fulls del llibre de treball. Al primer es fa una representació gràfica dels indicadors dels que es demana realitzar la projecció.

Perquè s'arribi al punt de col·lapse, la relació *NumTreb/NumPerceptors no actius* ha de ser dos. En aquest moment dos treballadors mantindran a un perceptor.

A la gràfica resultant de l'informe s'observa que aquesta relació no s'apropa a dos. Es veu que conforme passen els anys es va allunyant. Si es realitza una projecció d'aquest indicador, utilitzant la desviació respecte als anys anteriors, mai arribarà a dos.

Això passa amb les dades actuals que hi ha a la base de dades, però si les dades canvien es pot donar el cas d'arribar a aquest valor. És per això que es realitza el càlcul de quants anys han de transcórrer o han transcorregut per a que aquest indicador fos dos. Aquests càlculs estan explicats a l'apartat següent.


Projecció dels indicadors NumTreballadors/NumPerceptorsNoActius i NumTreballadors/Poblacio

Elementos de Página: Comunitat Autònoma: ANDALUCÍA

	Num Treballadors/Num Percep No Actius	Num Treballadors/Habitants
2005	1,42	0,41
2006	1,44	0,42
2007	1,41	0,43
2008	1,28	0,42
2009	1,15	0,40

Salaris Totals/Total Prestacions | Projecció Indicadors per de

Figura 29.- Informe per determinar el punt de col·lapse

A la figura anterior de manera numèrica i a la següent de manera gràfica, s'observa que l'indicador de col·lapse s'allunya de dos. La relació entre el número de treballadors i habitants es manté bastant lineal amb el pas dels anys.

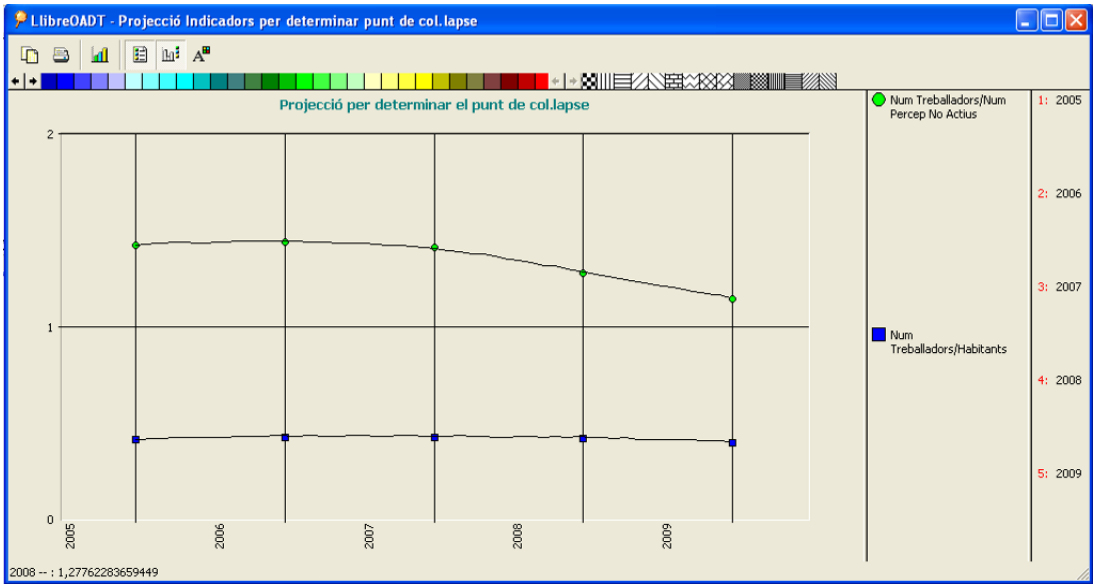


Figura 30.- Gràfic per determinar el punt de col·lapse

Càlcul de la projecció

Per realitzar el càlcul dels anys que han de passar (o han passat, llavors el valor resultant serà negatiu) es realitza una estimació tenint en compte la desviació entre els anys 2002 i 2009.

S'utilitza la fórmula:

2=Desviació del 2009+Número d'anys de distància a l'actual*Mitjana de la desviació

A on,

- El 2 indica que és el moment que dos treballadors mantenen un perceptor.
- La desviació del 2009 és la diferència entre Número de treballadors/Número de perceptors no actius del 2009 i el mateix indicador però del 2008.
- El número d'anys de distància és el número d'anys que han de passar o han passat perquè s'arribi al punt de col·lapse.
- La mitjana de la desviació és la mitjana de totes les desviacions calculades per tots els anys.

Per guardar els valors de les desviacions es crea una nova taula a la base de dades anomenada PROJECCIO. Aquesta taula conté els següents camps:

- IDANY: guarda l'identificador de l'any del que s'estan guardant les desviacions. Forma part de la clau primària, no pot ser NULL.
- IDCOMAUT: guarda l'identificador de la CA a la que pertanyen les dades. Forma part de la clau primària, no pot ser NULL.
- DESVIACIO: guarda la desviació respecte a l'any anterior.
- 2-DESVIACIO: es guarda el càlcul 2 menys desviació (2 és el valor que ens determina el punt de col·lapse), que és necessari per realitzar els càlculs.

El script de creació de la taula és el següent:

```
CREATE TABLE "PROJECCIO"
("IDANY" NUMBER(5,0) NOT NULL ENABLE,
 "IDCOMAUT" NUMBER(5,0) NOT NULL ENABLE,
 "DESVIACIO" NUMBER(5,3),
 "2-DESVIACIO" NUMBER(5,3),
  CONSTRAINT "PROJECCIO_PK" PRIMARY KEY ("IDANY", "IDCOMAUT") ENABLE   ) ;
```

Es crea un algorisme en PL/SQL per omplir automàticament la taula PROJECCIO. Es calcula la diferència de la relació *Número de treballadors/Número de perceptors no actius* i el mateix valor de l'any anterior. Això es fa en un bucle per cada any i CA. Quan es té aquest valor, s'omple el registre per inserir a la taula.

A continuació hi ha el pseudocodi de l'algorisme de càlcul dels valors i omplir la taula PROJECCIO.

```
//Es calculen i carreguen les dades de la desviació per posteriorment calcular el punt de col·lapse
//bucle principal per controlar que s'omplin tots els registres de desviació dels anys 2002-2009 i de totes les CCAA
Mentre no s'hagi arribat a l'any 2009
   //bucle per calcular la desviació dels registres de totes les comunitats d'un mateix any
   Mentre no s'hagin recorregut les 17 CCAA
         //es guarda la relació de treballadors/perceptors no actius de l'any en una variable
         Seleccionar actius/noactius i guardar a variable X de l'any i de la CA d'aquesta iteració
         //es guarda la relació de treballadors/perceptors no actius de l'any posterior en una variable
         Seleccionar actius/noactius i guardar a variable Y de l'any+1 i de la CA d'aquesta iteració
         //s'insereixen les dades de desviació calculades
         Inserir a la taula PROJECCIO (any+1, CA, Y-X, 2-(Y-X))
         Passar a següent CA
   Fi mentre
   Incrementar any
Fi mentre
Fi
```

Es pot consultar el script *CarregaProjeccio.sql* per veure el codi PL/SQL.

Un cop es tenen les desviacions de cada any respecte a l'anterior, es necessita saber la mitjana de les desviacions de tots els anys d'estudi per una mateixa CA. Es creu convenient la creació d'una vista materialitzada per mostrar aquest càlcul i poder realitzar les operacions de la fórmula final al Discoverer. La vista conté un camp amb la mitjana de les desviacions de tots els anys d'estudi i un identificador de la CA a la que pertanyen les dades.

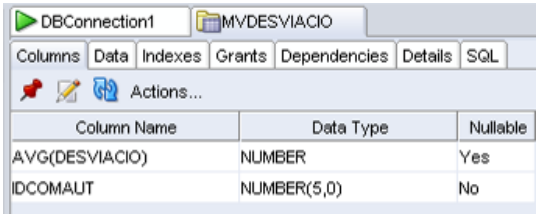


Figura 31.- Vista materialitzada MVDESVIACIO

El codi sql de la consulta de la creació de la vista materialitzada és el següent:

```
select avg(desviacio), idcomaut from projeccio, comunitatautonoma, temps where idany=temps.id and idComAut=comunitatautonoma.id group by idcomaut;
```

Cal crear dues carpetes noves a l'àrea de negoci i crear les unions necessàries per realitzar l'informe. Una carpeta per la taula PROJECCIO i una altra per la vista MVDESVIACIO.

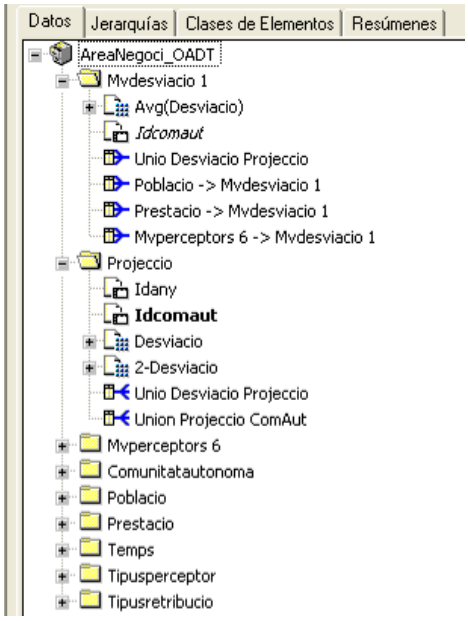


Figura 32.- Configuració definitiva de l'àrea de negoci

Un cop fet això ja es pot realitzar la consulta per saber quants anys han de passar per arribar al punt de col·lapse. Es pot comprovar a la següent imatge que tots els resultats són negatius, el que vol dir que no s'arribarà mai (segurament en anys anteriors si que s'hagués arribat). Això ja es podia intuir veient el gràfic amb les dades reals.

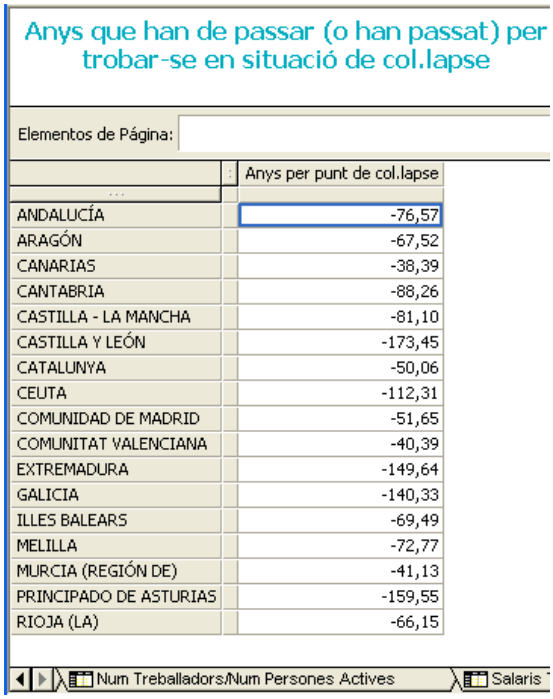


Figura 33.- Informe del anys per trobar-se al punt de col·lapse

Es realitza el càlcul a partir de la desviació obtinguda i de l'últim any del que es tenen dades reals.

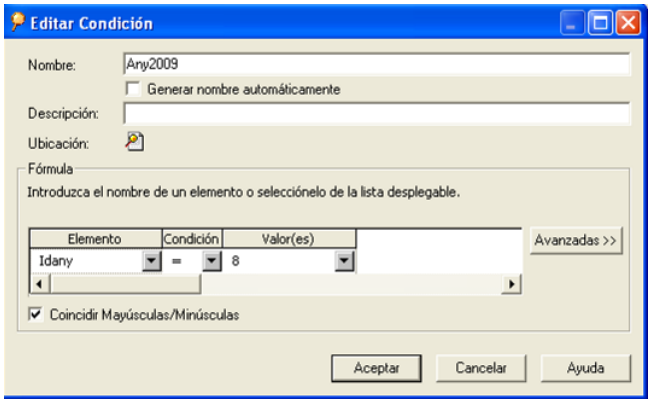


Figura 34.- S'imposa la condició que l'any sigui el 2009

El càlcul necessari en aquest informe es troba a continuació.

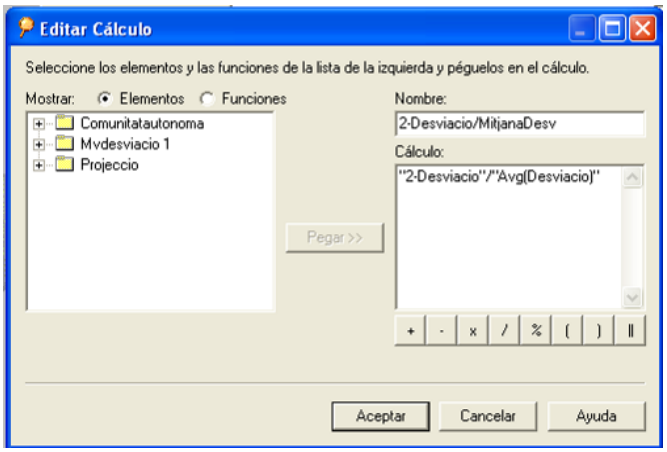


Figura 35.- Càlcul d'any de col·lapse

4. Conclusions

Amb la realització d'aquest projecte s'ha pogut veure el procés complet de creació d'un magatzem de dades, amb l'objectiu de poder realitzar determinades consultes i informes.

S'ha desenvolupat partint d'un anàlisi de requeriments, per posteriorment dissenyar i implementar el model multidimensional que possibilita la creació dels informes que el client requereix. Es carreguen les dades provinents de la font d'informació aportada, realitzant les tasques de manera automatitzada.

A més, s'ha realitzat la documentació i presentació relativa al producte final, explicant tots els processos seguits per la consecució del projecte.

Pel que a mi respecta, puc afirmar que, tot i les dificultats per les que he passat durant aquests mesos de treball, ha sigut una experiència molt positiva. He pogut posar en pràctica molts conceptes apressos durant els estudis a nivell més teòric. He adquirit coneixements d'un camp com és el dels magatzems de dades que per mi era desconegut, realitzant processos en llenguatge PL/SQL i utilitzant diferents eines amb les que no havia treballat com SqlDeveloper, Oracle Discoverer o SqlLoader. El fet de no tenir experiència amb aquestes eines ha suposat un esforç extra per desenvolupar el projecte, ja que he dedicat moltes hores aprenent el seu funcionament.

Es pot afirmar que s'han complert els objectius marcats a l'inici del projecte.

5. Línies d'evolució futura

El present projecte s'ha desenvolupat en base als requeriments del client i amb unes dades concretes. La solució aportada pot evolucionar i ampliar-se en diferents direccions, fent que siguin necessàries algunes modificacions a la implementació actual.

Pel que respecta a les dades emmagatzemades per aquest projecte, es té informació tributària que va des de l'any 2002 al 2009 i informació de població des de el 2005 al 2011. Perquè el magatzem de dades sigui útil per l'obtenció d'informació que serveixi per prendre decisions, cal que aquestes dades s'actualitzin periòdicament. Això vol dir que cada any s'han de carregar les dades proporcionades de l'any anterior. És necessari realitzar una sèrie de passos per poder introduir noves dades d'altres anys: modificació dels scripts de càrrega, afegir nous anys a la taula TEMPS, afegir noves columnes a algunes taules temporals on hi ha determinades columnes per cada any i modificar el codi PL/SQL que realitza els processos ETL (introduint funcions i parametritzant els anys).

És aconsellable que els fitxers font estiguin més unificats, que tots aportessin les mateixes dades dels mateixos anys. D'aquesta manera no seria necessari realitzar operacions per aproximar a uns valors que no s'han donat, tenint processos ETL més senzills i resultats de les consultes més realistes. Per això caldria informar a totes les CCAA perquè unifiquessin criteris.

El procés de càrrega de dades actualment és ràpid, però es té en compte que el volum de dades no és molt elevat. En cas que aquest volum creixés molt, seria interessant millorar els algorismes de càrrega a les taules definitives.

Pel que respecta a les consultes, els resultats s'obtenen ràpidament. En el disseny físic es té en compte la utilització d'índexs a les taules de la base de dades. Si amb un volum de dades molt superior es veu que el rendiment baixa, es poden afegir índexs segons les taules i les relacions que intervenen en cada informe.

Seria recomanable la utilització d'eines per l'obtenció dels informes via web, com Oracle Discoverer Plus. Així no seria necessària la instal·lació d'Oracle Discoverer Desktop a la part client i s'aportaria un entorn més senzill per l'usuari.

6. Glossari

Ad hoc (consulta): és una consulta a una base de dades que es pot personalitzar en temps real, en comptes de tenir una consulta dissenyada prèviament per obtenir un informe.

Agregació: és un procés de càlcul pel qual es resumeixen les dades dels registres de detall.

Atribut: és una característica amb la qual es pot descriure una entitat.

Taula de dimensió: és un element que conté atributs que s'utilitzen per restringir i agrupar les dades emmagatzemades en una taula de fets, quan es realitzen consultes sobre aquestes dades en un entorn de magatzem de dades.

Discoverer: és un conjunt d'eines de consulta ad hoc intuïtiva, anàlisis, informes i publicació en Web que proporciona als usuaris de negocis accés immediat a la informació de les bases de dades.

Discoverer Administrator: producte que serveix per crear, mantenir, administrar dades en la capa d'usuari final (EUL) i per definir com els usuaris interactuen amb les dades.

Discoverer Desktop: l'usuari final utilitza aquest component per executar consultes ad hoc i generar reports.

ETL: és el procés que permet a les organitzacions moure dades des de múltiples fonts, reformatar-les, netejar-les i carregar-les a una base de dades, data warehouse o a altre sistema operacional per ajudar a un procés de negoci.

Taula de Fets: és la taula central d'un esquema dimensional. Conté els valors de les mesures de negoci.

Magatzem de dades (Data Warehouse): és una col·lecció de dades orientada a un determinat àmbit, integrat, no volàtil i variable en el temps, que ajuda a prendre decisions a la entitat que l'utilitza.

OLAP: és una solució utilitzada en el camp de la intel·ligència empresarial (o Business Intelligence) que té l'objectiu d'agilitzar la consulta sobre grans quantitats de dades. Utilitza estructures multidimensionals que contenen les dades resumides de grans bases de dades o sistemes transaccionals. S'utilitza en informes de negocis de vendes, marketing, mineria de dades, etc.

Oracle: és un sistema de gestió de bases de dades objecte-relacional, desenvolupat per Oracle Corporation.

PL/SQL: és un llenguatge de programació incrustat a Oracle. Suporta totes les consultes, ja que la manipulació de dades és la mateixa que a SQL, incloent noves característiques.

Sistema de gestió de bases de dades: és un tipus de software molt específic, dedicat a servir d'interfície entre la base de dades, l'usuari i les aplicacions que l'utilitzen.

SQL (de l'anglés, Structured Query Language): és un llenguatge declaratiu d'accés a bases de dades relacionals que permet especificar diversos tipus d'operacions en aquestes. Permet efectuar consultes amb la fi de recuperar informació d'interès d'una base de dades, així com realitzar canvis sobre ella.

7. Bibliografia

Publicacions

Abelló Gamazo, Alberto (2003). "Disseny multidimensional". A: M. Serra Vizern i A. Rius Gavidia. *Magatzems de dades i models multidimensionals*. Barcelona : UOC. ISBN: 8484295818.

Enllaços d'Internet

Martínez Orol, Alfredo (03/2007). *OLAP y el diseño de cubos*. [en línia]. [data de consulta: 09/03/2012].
http://www.gestiopolis.com/canales8/ger/olap-online-analytic-processing.htm

"Arquitectura y concepto de un almacén de datos". *Etl tools info*. [en línia]. [data de consulta: 09/03/2012].
http://etl-tools.info/es/bi/almacendedatos_arquitectura.htm

"Snowflake Schema". *Data Warehousing*. [en línia]. [data de consulta: 14/03/2012].
http://www.1keydata.com/datawarehousing/snowflake-schema.html

WIKIPEDIA (2010). *Extract, Transform and Load*. [en línia]. [data de consulta: 07/04/2012].
http://es.wikipedia.org/wiki/Extract,_transform_and_load

WIKIPEDIA (2010). *PL/SQL*. [en línia]. [data de consulta: 08/04/2012].
http://es.wikipedia.org/wiki/PL/SQL

Calejero Román, Jesús Armand. *Índices en oracle.* [en línia]. [data de consulta: 17/04/2012].
http://www.slideshare.net/calejero/indices-en-oracle

FUJITSU ESPAÑA, SA. *Discoverer manual de usuario consultor.* [en línia]. [data de consulta: 03/05/2012].
http://www.gestion.uco.es/gestion/datawarehouse/doc/m_usuario/04000_consultor_discoverer.pdf

ORACLE (1996-2005). *Oracle® Business Intelligence Discoverer Desktop User's Guide*. [en línia]. [data de consulta: 06/05/2012].
http://download.oracle.com/docs/html/B13917_03/1intro.htm#sthref17

"Proyecciones con Oracle BI 10g What If". *Oracle BI Scorecards and Strategy Management*. [en línia]. [data de consulta: 09/05/2012].
http://blog.avanttic.com/2010/07/14/proyecciones-con-oracle-bi-10g-what-if/