Treball de Final de Carrera

Construcció i explotació d'un magatzem de dades per a l'anàlisi de vendes d'una cadena de supermercats

Memòria

Jordi Montané Balagué
Enginyeria Tècnica d'Informàtica de Gestió
Universitat Oberta de Catalunya
Curs 2012 -2013 - Primer Semestre

Consultor: Carles Llorach Rius

Data de lliurament: 7 de gener de 2013

“Tot gran viatge comença amb un pas”

Introducció i justificació del projecte

Com a treball de final de la carrera d'Enginyeria Tècnica d'Informàtica de Gestió s'ha de realitzar un treball de síntesi que permeti a l'estudiant desenvolupar un projecte dins d'un àrea proposada per la Universitat.

En la meva vida laboral i sobretot aquest últim any en que realitzava tasques de disseny de bases de dades i recopilació i explotació de dades en diversos formats m'he adonat que les organitzacions laborals treballen amb diversos formats de dades, tenen múltiples bases de dades no connectades que recullen informacions de diferents àrees d'interès, molts cops amb sistemes gestors diferents i d'altres vegades les dades arriben en arxius solts no relacionats amb cap eina de gestió de l'entitat.

Aquestes experiències en van fer veure la necessitat d'integrar les dades de manera eficient i poder donar informació útil i global als òrgans decisoris de l'organització i em vaig decidir per sol·licitar el meu TFC en l'Àrea de Magatzem de Dades.

Al demanar aquesta àrea de TFC em preguntava, és el Data Warehousing (com es coneix en anglès) la eina que permet atacar aquest problemes i donar-ne una solució efectiva?

Què és un Magatzem de dades?

És un conjunt de bases de dades i processos que integren dades de diferents fonts amb informació històrica i que té com a objectiu principal fer de suport en la presa de decisions. Aquesta funcionalitat s'aconsegueix mitjançant l'ús de diversos components informàtics.

- Unes fonts de dades operacionals.
- Una base de dades que té com a funció recopilar i consolidar les diferents dades operacionals. Com es veurà més endavant el disseny d'aquest base de dades difereix de l'habitual en les bases de dades relacionals. PER VEURE EL DISSENY ANAR A CAPÍTOL 2
- Processos d'extracció, transformació i càrrega de dades -els processos ETL (sigles en anglès) Mitjançant aquest processos es traspassen les dades dels diferents orígens a la base de dades del magatzem. PER VEURE EL PROCÉS ETL ANAR A CAPÍTOL 3
- Processos de producció d'informació en forma d'informes resum de dades i gràfics Aquestes eines proporcionen la informació processada i consolidada.

En els capítols d'aquesta memòria que segueixen a continuació exposo els passos fets i els procediments i coneixements emprats en la realització del projecte.

Índex de continguts

Capítol 1 - Planificació del projecte

I - Objectius del TFC

El treball de final de carrera com a part final d'uns estudis té objectius formatius de síntesi dels coneixements estudiats al pla d'estudis enfocats cap a l'obtenció d'un producte funcional.

Objectius Generals

l'Objectiu principal del projecte és dissenyar, construir i explotar un magatzem de dades a partir de la informació disponible en orígens de dades diversos, bases de dades transaccionals i arxius de full de càlcul o de tipus CSV.

Treballar les tècniques de gestió de projectes i anàlisi de requeriments tot i seguint les necessitats d'un client.

Objectius Específics

Construir un magatzem de dades: adquirir experiència en el disseny i implementació d'una base de dades per a la consolidació de dades d'explotació obtingudes de diverses fonts i formes.

Treballar les tècniques d'extracció, transformació i càrrega de dades (ETL)

Conèixer i practicar les tècniques d'explotació de dades, realitzar informes de dades.

Produir la memòria d'un projecte que n'expliqui tant la gestació com la seva funció.

II - Requeriments de la solució informàtica

Precedents

En l'enunciat proposat del TFC es demana la **construcció i explotació d'un magatzem de dades per a l'anàlisi de vendes d'una cadena de supermercats,** la corporació Grup Líder en Distribució de Proximitat (GLDP)

L'empresa GLDP ha percebut la situació de crisi econòmica actual i pateix una disminució de vendes amb la qual cosa vol estudiar la seva xarxa d'establiments per tal de millorar-ne el rendiment.

El problema que observen els òrgans decisors de GLDP és que degut a la gestió distribuïda de les compres de cadascun dels establiments del grup i atès que la consolidació de dades de vendes és trimestral, no tenen accés a les dades en temps i forma adients per a avaluar-les per tal d'engegar les campanyes de màrqueting i l'obtenció de millors resultats en els establiments i períodes més fluixos.

Ens han encarregat la creació d'una solució informàtica amb una sèrie de requisits que s'analitzen al següent apartat.

III - Anàlisi de requisits

Requeriments funcionals

Model de domini de l'aplicació

De la documentació de l'encàrrec del Grup Líder en Distribució de Proximitat (GLDP) en podem extreure la següent informació.

Els establiments del grup trameten un arxiu *nomestabliment.accdb* amb la relació de vendes, GLDP proporciona arxius de tipus CSV amb el catàleg de productes de l'any, *productesXXX.csv*, i també un arxiu *Establiments.xls* amb les dades sobre els establiments que formen el grup de distribució GLDP.

Un usuari ha de fer la càrrega de dades: recopilar les dades rebudes, normalitzar-les i incorporar-les al magatzem de dades.

El magatzem proporcionarà informació sobre les dades mostrant-les amb agregació per demarcació territorial, tipus d'establiment i família de productes i amb temporalitat mensual i anual.

Podem identificar els objectes Establiment, GLDP, nomestabliment.accdb, productes20XX.csv, establiments.xls, càrrega de dades, informes, usuari, agregació de dades.

Identificació dels actors

Els actors a definir son dos, l'usuari **administrador** i el **consultor**. No son actors els establiments, atès que no interactuen amb el magatzem de dades directament, sinó que només hi proporcionen dades.

Ambdós actors son dependents doncs es defineix que administrador pot executar tots els casos d'ús de consultor a més dels seus exclusius. Administrador per tant és una especialització de consultor.

Consultor és un usuari final, que té tasques d'accés a dades, impressió de dades, però no té tasques de gestió del sistema magatzem de dades. Administrador és un usuari del sistema i pot gestionar els usuaris del sistema, carregar noves dades, crear mecanismes ETL i gestionar la base de dades.

Identificació dels casos d'ús

Casos d'ús exclusius de l'administrador

1. Extracció de dades (ETL)
2. Normalització de dades (ETL)
3. Càrrega de dades (ETL)
4. Gestió d'usuaris del magatzem de dades
5. Gestió base de dades del magatzem de dades
6. Crear nova consulta de selecció
7. Crear nou informe

Casos d'ús de consultor (i per especialització també d'administrador)

8. Consulta i impressió d'informes de dades : Informe de total de vendes i marge net del grup, Import mitjà de compra per establiment, % de vendes respecte el total del grup (per establiment i per demarcació), Rànking d'establiments per nombre de vendes i volum total, "Top ten" de productes...

Requeriments no funcionals.

Restriccions de l'entorn de dades

En la recepció de dades es d'esperar que el format de les dades rebudes en el cas de fulls d'Excel (Establiments.xls) o CSV (Productes anyXXXX) no segueixi una pauta constant durant el temps. És conegut que els usuaris de fulls de càlcul acostumen a variar tant l'estructura dels fulls com la especificació de les dades.

Així doncs el model ETL programat té un punt dèbil: l'alta probabilitat d'alteració de les fonts de dades a tractar. Atesa aquesta situació cal establir un mecanisme de revisió de les fonts Establiments.xls i Productes anyXXXX rebudes.

Cal que l'usuari encarregat de fer la càrrega de dades, executi un protocol de revisió dels arxius rebuts.

- Preservar l'arxiu rebut amb datació del moment de rebuda i treballar amb una còpia del mateix.
- Comprovar que l'estructura de les dades rebudes sigui conforme al model ETL establert, en cas contrari caldrà requerir un nou arxiu o redefinir el model ETL per a preservar l'equivalència correcta entre els camps de dades origen en els arxius rebuts i els camps destí del magatzem de dades.

L'actor administrador ha de ser l'operador que executi aquesta tasca atès que també és el que executa la càrrega de les dades.

En el cas de dades rebudes en format base de dades Access (dades de vendes de cada establiment) els establiments del grup usen una base de dades estàndard per a la recollida de dades de vendes.

Pel que fa les dades de població necessàries per als informes, s'obtindran del portal d'internet IDESCAT i es farà la mitjana entre la dada de 1r de gener i la de 31 de desembre.

VEURE COMENTARIS SOBRE LES DADES EN L'APARTAT DE COMENTARIS DE LA FASE D'IMPLEMENTACIÓ

Requisits de manteniment i extensibilitat

Tot i assegurar el màxim possible pels procediments anteriors la fiabilitat del model ETL, cal tenir en compte que es produiran amb el pas del temps nous requeriments d'informació i també canvis en els orígens de dades (arxius MDB, CVS, Excel) per tant, el model ETL requerirà d'actualitzacions i manteniment. Si GLDP no té capacitat per fer l'actualització i manteniment cal establir amb el client un contracte de manteniment .

IV - Organització del projecte

Per a construir el magatzem de dades es divideix el projecte en tres fases

Fase 1 - Pla de Treball	→ Planificació de la feina, previsió d'activitats i terminis
Fase 2 - Anàlisi i Disseny	→ Anàlisi detallat de requeriments, disseny del model dimensional i ETL
Fase 3 - Implementació	→ Programació de la base de dades, de les eines ETL i Informes

Components de Software i Hardware

Pel desenvolupador del TFC

Programari - Software

Sistema operatiu: Windows 7 professional 64bits Service Pack 1
Diagrames amb MS Visio 2010 i LucidChart
Eina Planificació MS Project 2010
Eines de documentació i tractament de dades MS Office 2007 → Word, Excel i Access
Instal·lació de màquina virtual de VirtualBox de l'entorn proporcionat per la UOC

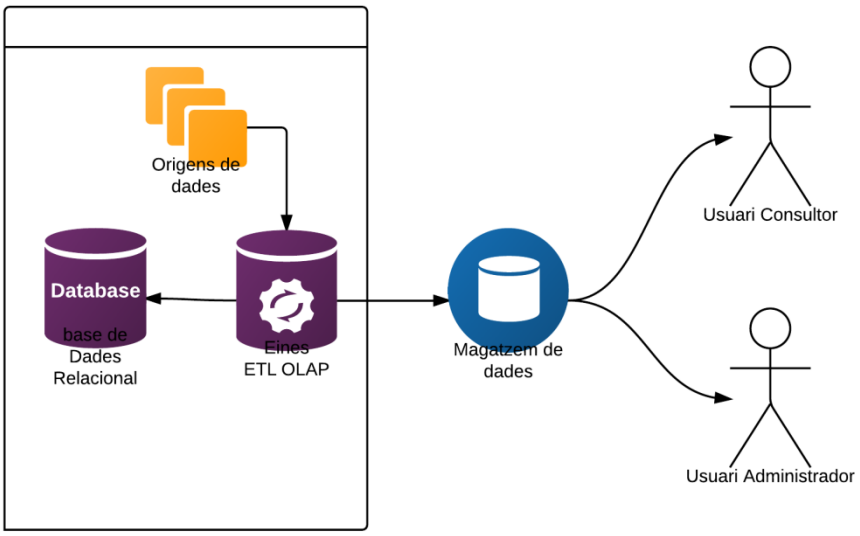
Maquinari - Hardware

PC amb 4 Processadors duals Intel Core i7 a 2.80GHz, RAM 8GB, disc dur ATA de 1TB, disc extern USB 500Mb

Pel client

Atès que l'enunciat no especifica la situació tecnològica del client, estableixo com a supòsit de treball que el client GLDP té una infraestructura de hardware suficient per a la instal·lació de la solució informàtica i que té també les llicències de software necessàries per a tenir el magatzem de dades en les tecnologies de la solució informàtica proposada.

Arquitectura del projecte



Anàlisi de riscos

En la següent taula es recullen el tipus d'incidències previsibles en el desenvolupament del TFC i la solució proposada en cas de aquestes que es produeixin.

Incidència	Pla de contingència
Avaria del Punt de treball	Replicar l'entorn de treball i fer còpies de seguretat en disc dur extern amb periodicitat. Copia incremental diària i còpia completa setmanal. Substituir el maquinari avariat
Malaltia	Atès que els recursos de personal es limiten a l'estudiant del TFC, en cas de malaltia lleu no se suspendrà cap activitat. En cas de malaltia amb conseqüències d'incapacitat temporal llarga caldrà suspendre el desenvolupament del TFC i prorrogar-lo al següent semestre.

V - Proposta d'activitats i cronograma

Relació d'activitats, Estimació de temps i Fites a complir

# Tasca	Activitats a realitzar	Inici	Fi	Prede-cessores	Notes
1	**TFC - Magatzem de dades**	**20/09/12**	**07/01/13**		
2	**Fase 1 - Inici TFC**	**20/09/12**	**02/10/12**		
3	Obtenir recursos del web de la UOC: documentació i materials de l'assignatura	23/09/12	23/09/12		Manca el software de la màquina virtual per a fer el desenvolupament del magatzem virtual. Caldrà obtenir-lo i instal·lar-lo més endavant. Com a molt tard a l'inici de la fase d'implementació

# Tasca	Activitats a realitzar	Inici	Fi	Prede-cessores	Notes
4	Revisió de la documentació i software	24/09/12	24/09/12	3	
5	Instal·lació software de suport	24/09/12	24/09/12	4	
6	Trobada inicial amb el consultor	25/09/12	25/09/12		
7	**Fase 2 - Pla de treball TFC - PAC1**	**26/09/12**	**02/10/12**		
8	Revisió del cas proposat a l'enunciat	26/09/12	26/09/12		
9	Elaboració del Pla de treball	26/09/12	02/10/12		
10	Anàlisi preliminar de requeriments de la solució informàtica	26/09/12	28/09/12	8	
11	Finalització i Correcció PAC 1	01/10/12	02/10/12		
12	Fita 1 - Lliurament PAC1	02/10/12	02/10/12		
13	**Fase 3 - Anàlisi i Disseny de la solució informàtica - PAC2**	**03/10/12**	**06/11/12**		
14	Estudi conceptes nous relacionats amb DW: eines ETL, OLAP	03/10/12	12/10/12		
15	Anàlisi de requeriments de la solució informàtica	09/10/12	13/10/12		
16	Disseny de la base de dades	14/10/12	22/10/12	15	
17	Disseny de les eines ETL i OLAP	23/10/12	29/10/12	16	
18	Disseny informes	30/10/12	03/11/12	17	
19	Instal·lació de Màquina Virtual del projecte	02/11/12	03/11/12		
20	Revisió PAC2	04/11/12	06/11/12		
21	Fita 2 - Lliurament PAC2	06/11/12	06/11/12	20	
22	**Fase 4 - Implementació del magatzem de dades - PAC3**	**07/11/12**	**19/12/12**		
23	Creació física de la base de dades	07/11/12	22/11/12		
24	Implementació dels processos ETL	23/11/12	30/11/12	23	
25	Implementació dels processos OLAP	01/12/12	08/12/12	24	
26	Implementació dels informes	09/12/12	16/12/12	25	
27	Revisió PAC3	17/12/12	19/12/12		
28	Fita 3 - Lliurament PAC3	19/12/12	19/12/12	27	
29	**Fase 5 - Lliurament final. Memòria i presentació virtual**	**20/12/12**	**07/01/13**		
30	Elaboració de la memòria final	20/12/12	07/01/13		

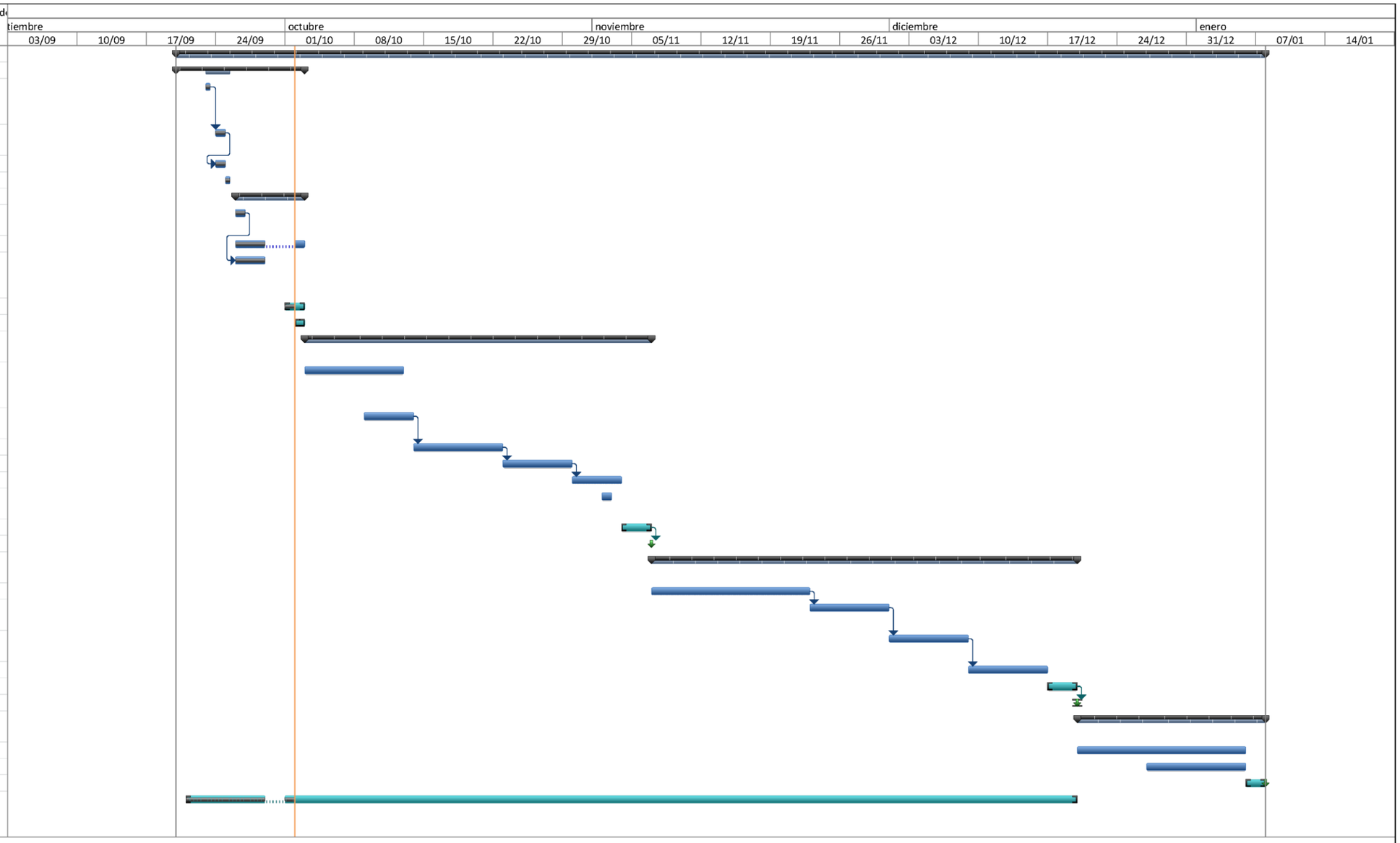
# Tasca	Activitats a realitzar	Inici	Fi	Prede-cessores	Notes
31	Elaboració presentació virtual	27/12/12	07/01/13		
32	Fita 4 - Revisió i Lliurament Final	06/1/13	07/1/13		
33	Recopilació de dades i documentació per a la memòria del projecte	21/09/12	07/01/13		

Resum de les fites a complir per a completar el TFC

Fita 1 - Lliurament PAC1	Pla de treball	02/10/12
Fita 2 - Lliurament PAC2	Anàlisi i Disseny de la solució informàtica	06/11/12
Fita 3 - Lliurament PAC3	Implementació del magatzem de dades	19/12/12
Fita 4 - Lliurament Final	Memòria i presentació virtual	07/01/13

Diagrama de Gantt

Id	Nombre de tarea	Comienzo	Fin	Pred
1	**TFC - Magatzem de dades**	**jue 20/09/12**	**lun 07/01/13**	
2	**Fase 1 - Inici TFC**	**jue 20/09/12**	**mar 02/10/12**	
3	Obtenir recursos del web de la UOC: documentació i materials de l'assignatura	dom 23/09/12	dom 23/09/12	
4	Revisió de la documentació i software	lun 24/09/12	lun 24/09/12	3
5	Instal·lació software de suport	lun 24/09/12	lun 24/09/12	4
6	Trobada inicial amb el consultor	mar 25/09/12	mar 25/09/12	
7	**Fase 2 - Pla de treball TFC - PAC1**	**mié 26/09/12**	**mar 02/10/12**	
8	Revisió del cas proposat a l'enunciat	mié 26/09/12	mié 26/09/12	
9	Elaboració del Pla de treball	mié 26/09/12	mar 02/10/12	
10	Anàlisi preliminar de requeriments de la solució informàtica	mié 26/09/12	vie 28/09/12	8
11	Finalització i Correcció PAC 1	lun 01/10/12	mar 02/10/12	
12	Fita 1 - Lliurament PAC1	mar 02/10/12	mar 02/10/12	
13	**Fase 3 - Anàlisi i Disseny de la solució informàtica - PAC2**	**mié 03/10/12**	**mar 06/11/12**	
14	Estudi conceptes nous relacionats amb DW: eines ETL, OLAP	mié 03/10/12	vie 12/10/12	
15	Anàlisi de requeriments de la solució informàtica	mar 09/10/12	sáb 13/10/12	
16	Disseny de la base de dades	dom 14/10/12	lun 22/10/12	15
17	Disseny de les eines ETL i OLAP	mar 23/10/12	lun 29/10/12	16
18	Disseny informes	mar 30/10/12	sáb 03/11/12	17
19	Instal·lació de Màquina Virtual del projecte	vie 02/11/12	sáb 03/11/12	
20	Revisió PAC2	dom 04/11/12	mar 06/11/12	
21	Fita 2 - Lliurament PAC2	mar 06/11/12	mar 06/11/12	20
22	**Fase 4 - Implementació del magatzem de dades - PAC3**	**mié 07/11/12**	**mié 19/12/12**	
23	Creació física de la base de dades	mié 07/11/12	jue 22/11/12	
24	Implementació dels processos ETL	vie 23/11/12	vie 30/11/12	23
25	Implementació dels processos OLAP	sáb 01/12/12	sáb 08/12/12	24
26	Implementació dels informes	dom 09/12/12	dom 16/12/12	25
27	Revisió PAC3	lun 17/12/12	mié 19/12/12	
28	Fita 3 - Lliurament PAC3	mié 19/12/12	mié 19/12/12	27
29	**Fase 5 - Lliurament final. Memòria i presentació virtual**	**jue 20/12/12**	**lun 07/01/13**	
30	Elaboració de la memòria final	jue 20/12/12	sáb 05/01/13	
31	Elaboració presentació virtual	jue 27/12/12	sáb 05/01/13	
32	Fita 4 - Revisió i Lliurament Final	dom 06/01/13	lun 07/01/13	
33	Recopilació de dades i documentació per a la memòria del projecte	vie 21/09/12	mié 19/12/12	



Proyecto: Projecte GLDP - Pla de T
Fecha: mar 02/10/12

Tarea · Resumen · Hito externo · Resumen inactivo · Informe de resumen manual · Sólo fin
División · Resumen del proyecto · Tarea inactiva · Tarea manual · Resumen manual · Fecha límite
Hito · Tareas externas · Hito inactivo · Sólo duración · Sólo el comienzo · Progreso

Página 1

Capítol 2 - Anàlisi i disseny

En aquest capítol es descriu el procés d'anàlisi de les dades i s'estudia el magatzem de dades, des del seu model conceptual fins a la definició del model físic de la base de dades que contindrà les dades definitives.

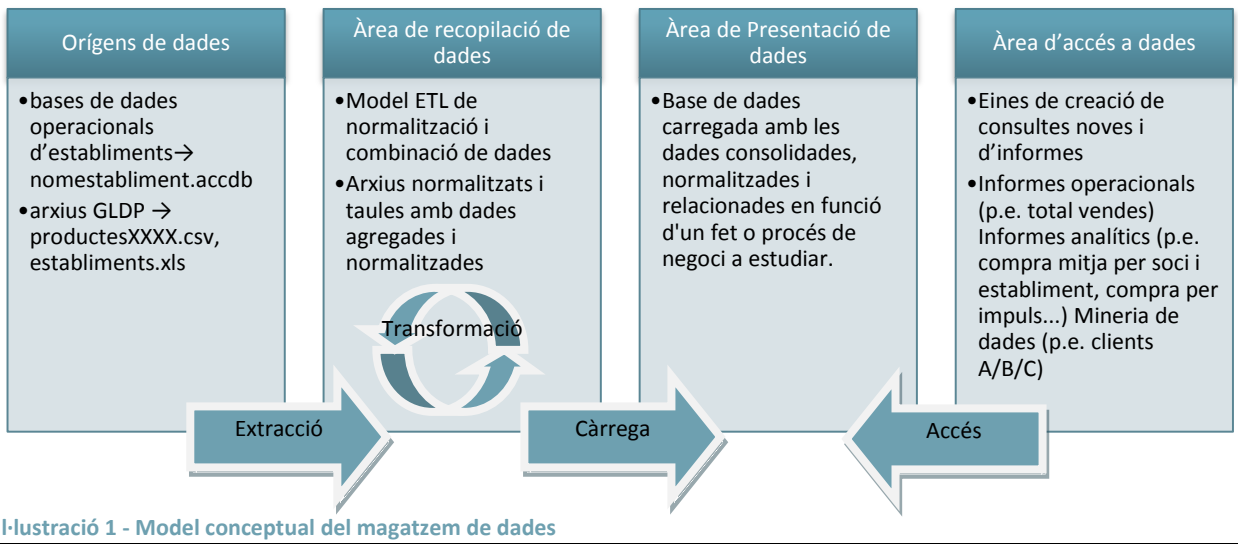
I - Model Conceptual del magatzem de dades

A partir de la definició de magatzem de dades se'n dedueixen quatre conceptes o àrees constituents: l'àrea dels orígens de dades; l'àrea de recopilació, transformació i càrrega de dades; l'àrea de presentació de dades definitives i l'àrea d'accés a dades.

A cadascuna d'aquestes àrees hi podem atribuir uns objectes físics (arxius informàtics) i uns processos informàtics que les caracteritzen.

- A l'àrea d'orígens de dades hi trobarem com a objectes físics, els diversos arxius que son font de dades, aquests arxius poden provenir tant de la pròpia empresa com de recerques de dades a Internet.
- A l'àrea de recopilació de dades hi trobarem la especificació dels processos d'extracció, de transformació i de càrrega de dades i també arxius de dades ja transformades i normalitzades.
- A l'àrea de presentació de dades hi trobarem la base de dades amb les dades definitives carregades.
- A l'àrea d'accés a dades hi trobarem les eines per visualitzar les dades i els arxius d'informes definits.

La il·lustració següent resumeix el model conceptual d'un magatzem de dades amb els objectes i processos per a aquest projecte.



Il·lustració 1 - Model conceptual del magatzem de dades

II - Disseny de la Base de Dades - Diagrama E-R

Per tal de generar una base de dades cal estudiar els tipus de dades que contindrà, la informació que se'n vol deduir i els orígens de dades disponibles.

A continuació es presenta l'anàlisi dels arxius origen de dades per tal de dissenyar una primera estructura de base de dades relacional en forma de diagrama Entitat - Relació.

Relació d'orígens de dades del projecte GLDP

Nom arxiu	Contingut	Tipus d'arxiu
Productes 2010.csv	catàleg de productes del grup de l'any 2010	Arxiu de registres de dades separades per marcador
Productes 2011.csv	catàleg de productes del grup de l'any 2011	
Productes 2012.csv	catàleg de productes del grup de l'any 2012	
Establiments.xlsx	relació d'establiments del grup GLDP	Arxiu de dades en full de càlcul Excel 2007
Figueres.accdb	dades de vendes de l'establiment Figueres GLDP	Arxiu de base de dades de MS Access 2007
Girona.accdb	dades de vendes de l'establiment Girona GLDP	
Lleida.accdb	dades de vendes de l'establiment Lleida GLDP	
Olot.accdb	dades de vendes de l'establiment Olot GLDP	
Poble Sec.accdb	dades de vendes de l'establiment Poble Sec GLDP	
Reus.accdb	dades de vendes de l'establiment Reus GLDP	
Sants.accdb	dades de vendes de l'establiment Sants GLDP	
Tarragona.accdb	dades de vendes de l'establiment Tarragona GLDP	
Terrassa.accdb	dades de vendes de l'establiment Terrassa GLDP	
Tortosa.accdb	dades de vendes de l'establiment Tortosa GLDP	
Vielha.accdb	dades de vendes de l'establiment Vielha GLDP	

Dels orígens de dades en podem deduir les següents entitats i atributs.

Origen de dades → Establiments.xls

Establiment (id_establiment, nom, superfície, dia_soci, cost_fix_mensual, tipus)
On id_establiment és clau principal

Origen de dades → Productes 20XX.csv

Producte (id_producte, nom_producte, preu_venda, proveidor, tipus_producte, IVA, preu_cost, codi_barres, venda_impuls)
On id_producte és clau principal

Origen de dades → Figueres.accdb, Girona.accdb, …, Vielha.accdb

Venda(id_venda, data, hora, forma_pagament, id_client, descompte)
On id_venda és clau principal
On id_client és clau forana

Detall_venda(id_venda, id_linia, unitats, id_producte)

On id_venda, id_linia,id_producte son claus foranes

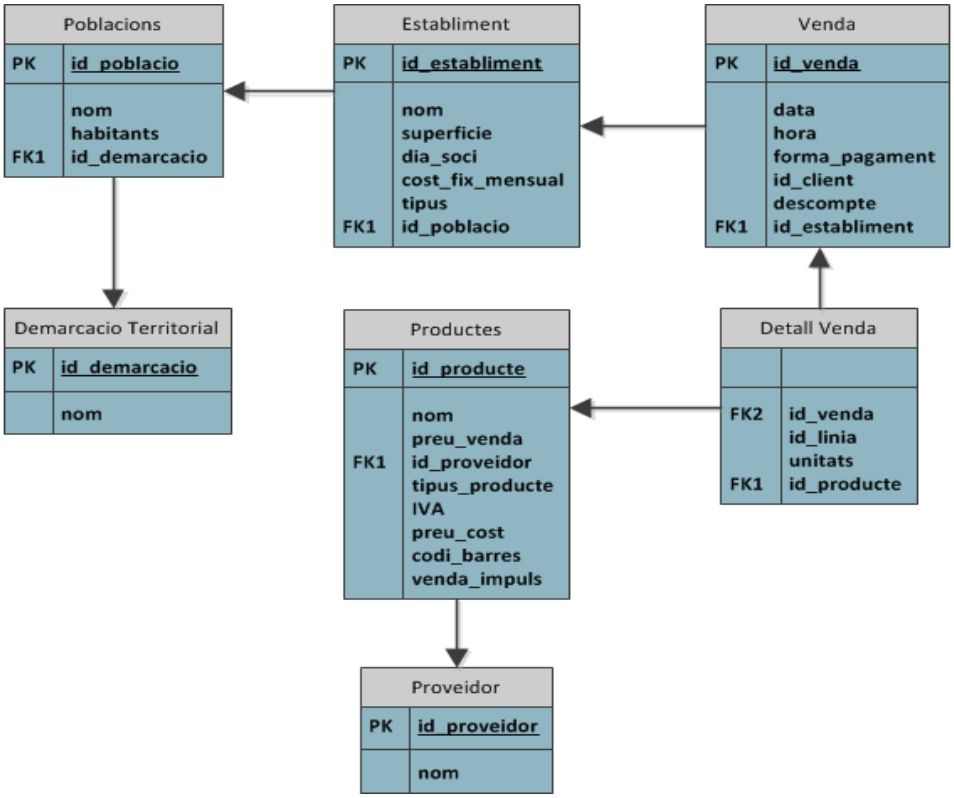
Dels requisits funcionals deduïm que:

Per a poder fer les agregacions per demarcació territorial necessitem afegir l'entitat **Demarcacio_territorial.**

Per a estudiar les dades segons la població cal crear l'entitat **Poblacions** i relacionar-la amb l'entitat **Establiment** i també establir la relació forana id_demarcacio a l'entitat **Poblacions**.

Per a fer les agregacions de vendes per establiment cal relacionar les entitats **Establiment** i **Venda** afegint l'atribut id_establiment a aquesta última entitat.

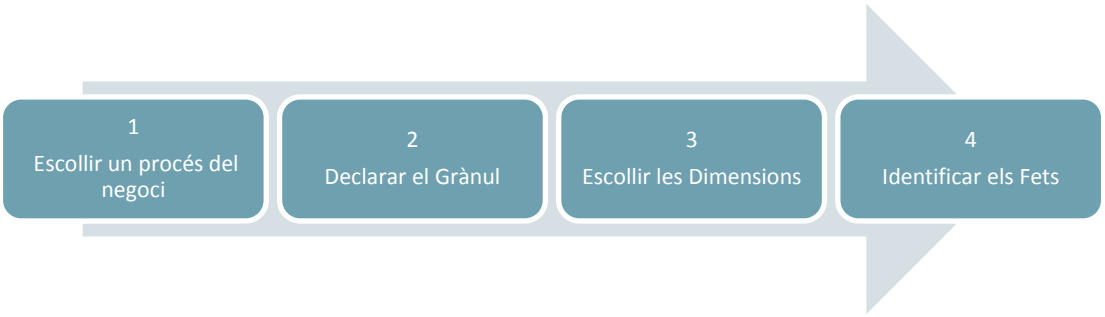
A partir de l'anàlisi de dades construeixo el següent diagrama entitat relació



III - Model multi-dimensional detallat

Com comentava en la introducció hi ha diferències en el tractament de les entitats i de les relacions quan es tracta de construir un magatzems de dades i no una base de dades convencional. En aquest projecte he seguit la teoria explicada per R.Kimball i M. Ross en el llibre The Data Warehouse Toolkit considerant també el model d'entitat relació obtingut en la fase prèvia d'anàlisi.

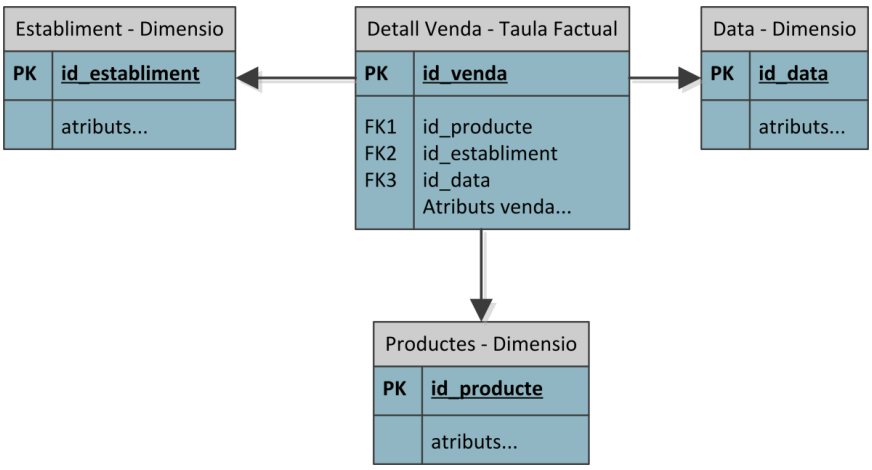
Segons els autors per a produir el model dimensional cal seguir un procés en 4 fases



Punt 1. De l'encàrrec del projecte de GLDP se'n desprèn que l'objecte d'estudi principal del magatzem de dades ha de ser el procés que doni resposta a la qüestió plantejada a l'enunciat *"conèixer millor el comportament de les vendes per tal de capgirar-ne l'evolució negativa dels darrers anys"* . Escullo el procés de venda com a procés de negoci a estudiar.

Punt 2. Cal escollir el gràgul de manera que aporti la informació més detallada sobre el procés a estudiar, la dada més atòmica (indivisible) possible, en el nostre cas son les vendes al detall. El gràgul ens identifica quina serà la taula factual.

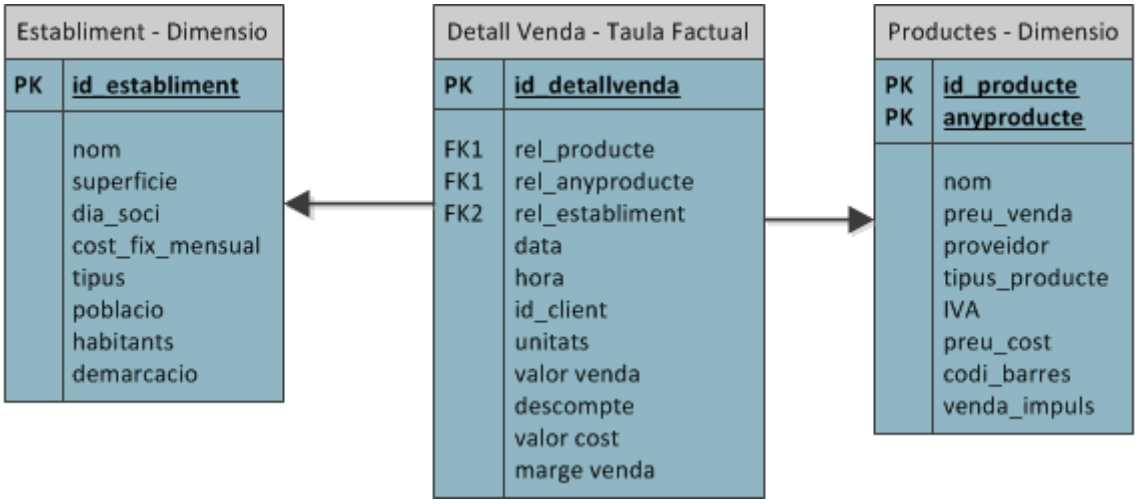
Punt 3. Les dimensions les escollim en funció de com es descriuen les dades atòmiques dins el procés del negoci escollit. La data de la venda, els productes venuts i l'establiment on s'ha produït la venda son les dimensions que cal estudiar. En aquest punt ja podem visualitzar un esquema bàsic del model dimensional. Mes endavant poblarem els atributs de cadascuna de les dimensions a partir dels atributs detectats als orígens de dades.



Punt 4. En aquest punt hem d'escollir els fets que apareixeran a la taula factual, **hem d'escollir els atributs que ens permetin explorar i explotar adequadament les dades numèriques de la venda per a satisfer els requisits funcionals**. Ens interessa que el detall d'una venda inclogui la quantitat d'aquest producte venut, el valor total de la venda, el cost, el valor total del descompte i el marge de benefici. Totes aquestes dades les obtenim dels orígens de dades.

Atès el requisit "*% de marques blanques per habitant"* cal afegir l'atribut nombre d'habitants de la població a la dimensió Establiment. Per a poder classificar els clients A/B/C cal mantenir l'atribut "*client"* a la taula factual encara que no sigui un valor additiu entre registres

Model dimensional detallat un cop revisades les dimensions i els fets.



Nota: En fase d'implementació he simplificat la dimensió data i he atribuït els valors a la taula factual.

Un cop revisats els orígens de dades i dissenyat el model dimensional de la base de dades del magatzem, passem al següent capítol per estudiar i programar els processos d'extracció, transformació i càrrega de dades.

Capítol 3 - Implementació ETL i Informes

En aquest capítol es descriuen detalladament els processos d'extracció, transformació i càrrega de dades per a cada tipus d'origen de dades i es mostren els informes de dades realitzats.

I - Què son els processos ETL?

Com em vist anteriorment un dels components d'un magatzem de dades és el procés ETL. Aquest procés consisteix en aplicar als orígens de dades tres procediments (o processos)que s'executen de manera consecutiva, el procés d'extracció, el de transformació i el de càrrega.

L'objectiu final del procés complet ETL és obtenir a partir dels orígens de dades el conjunt de dades necessàries i suficients convenientment normalitzades per a poder estudiar un procés de negoci determinat pels requeriments.

El procés d'extracció consisteix en obtenir les dades d'origen. Es poden donar diferents casos: rebre un fitxer amb les dades, pot ser que es rebi una base de dades sencera i que sigui necessari executar sentencies SQL per obtenir les dades significatives, o pot ser que calgui cercar les dades a fonts externes de manera manual.

El procés de transformació consisteix en aplicar a les dades obtingudes diverses operacions per tal d'obtenir dades normalitzades i útils per als objectius del magatzem de dades. Aquestes operacions s'efectuen mitjançant sentències SQL amb programes de gestió de bases de dades, o amb funcions de gestors de fulls de càlcul. Alguns exemples d'operacions a fer son: normalitzar identificadors, assegurar valors vàlids en cada categoria, evitar valors nuls, també es produeixen dades calculades a partir de les diverses fonts segons les necessitats del projecte p.e. preus calculats de la venda.

El procés de càrrega consisteix en executar sentencies d'inserció de dades des de la base de dades final del magatzem per inserir-hi les dades processades anteriorment.

Periodicitat d'extracció de dades

El procés d'extracció pot ser únic, o pot produir-se repetidament durant l'any natural. També cal distingir entre els casos en que l'extracció serà complerta, o es faran extraccions incrementals, pe. les dades de vendes seran des d'inici de període fins a data d'extracció, o seran només les del mes finalitzat?

Per a aquest projecte s'han identificat com a orígens de dades: els arxius de vendes de cada establiment en format de base de dades, els arxius de productes anuals i l'arxiu de llistat d'establiments.

L'enunciat no especifica la regularitat d'actualització de dades, tot i això, a continuació n'especifico la periodicitat d'extracció i actualització aplicant criteris habituals temporals segons escaigui:

- Per a les dades de llistat d'establiments (arxiu establiments.xlsx) s'estableix una extracció anual, i actualització només en cas de notificació específica de canvis en les dades del establiments.

- Per a les dades de productes (arxius productes 20XX.csv) estableixo una càrrega inicial de tots els arxius i en cas que el grup GLDP comuniqui canvis en el catàleg de productes, es farà una actualització mensual incremental.

- Per als arxius de vendes dels establiments (arxius nomestabliment.accdb) estableixo també una càrrega inicial total i l'actualització mensual incremental a mes tancat en funció de la tramesa de dades dels establiments.

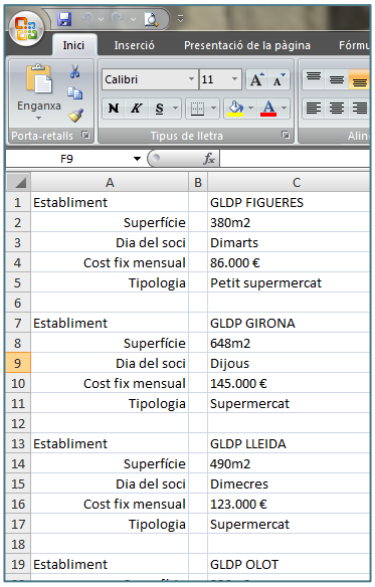
II - Processos ETL segons l'origen de dades

Els següents apartats descriuen els diferents processos realitats d'extracció, normalització (o transformació) i càrrega de dades per a cada tipus d'arxiu origen de dades

Origen: arxiu establiments.xls

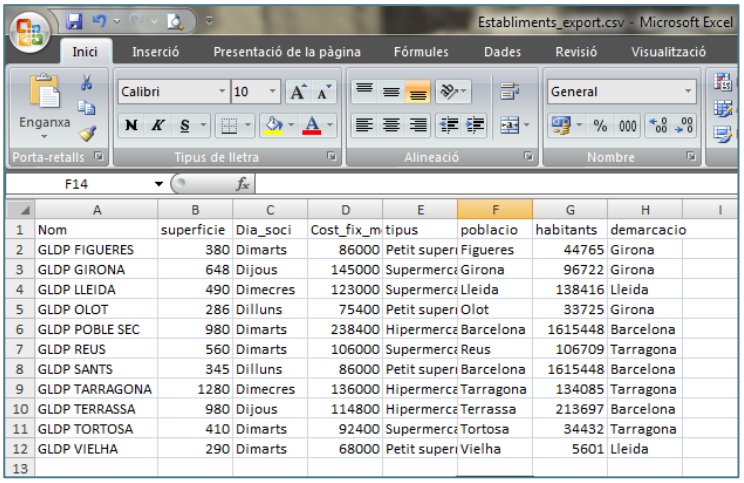
Extracció → Transposició de valors de l'arxiu Excel distribuïts en forma de fitxa a format taula

Transformació → Neteja de valors numèrics (treure formats de moneda al camp cost fix mensual, qualificador 'm2' al camp superficie) inserció de camps i valors de població, habitants i demarcació. Els valors d'habitants i demarcació s'han obtingut del portal web IDESCAT que les proporciona actualitzades a 01.01.2011. Aquests canvis s'han executat des de l'aplicació Excel.



	A	B	C
1	Establiment		GLDP FIGUERES
2	Superfície		380m2
3	Dia del soci		Dimarts
4	Cost fix mensual		86.000 €
5	Tipologia		Petit supermercat
6			
7	Establiment		GLDP GIRONA
8	Superfície		648m2
9	Dia del soci		Dijous
10	Cost fix mensual		145.000 €
11	Tipologia		Supermercat
12			
13	Establiment		GLDP LLEIDA
14	Superfície		490m2
15	Dia del soci		Dimecres
16	Cost fix mensual		123.000 €
17	Tipologia		Supermercat
18			
19	Establiment		GLDP OLOT

Il·lustració 2 - Estructura Original



	A	B	C	D	E	F	G	H	I
1	Nom	superficie	Dia_soci	Cost_fix_m	tipus	poblacio	habitants	demarcacio	
2	GLDP FIGUERES	380	Dimarts	86000	Petit superi	Figueres	44765	Girona	
3	GLDP GIRONA	648	Dijous	145000	Supermerca	Girona	96722	Girona	
4	GLDP LLEIDA	490	Dimecres	123000	Supermerca	Lleida	138416	Lleida	
5	GLDP OLOT	286	Dilluns	75400	Petit superi	Olot	33725	Girona	
6	GLDP POBLE SEC	980	Dimarts	238400	Hipermerca	Barcelona	1615448	Barcelona	
7	GLDP REUS	560	Dimarts	106000	Supermerca	Reus	106709	Tarragona	
8	GLDP SANTS	345	Dilluns	86000	Petit superi	Barcelona	1615448	Barcelona	
9	GLDP TARRAGONA	1280	Dimecres	136000	Hipermerca	Tarragona	134085	Tarragona	
10	GLDP TERRASSA	980	Dijous	114800	Hipermerca	Terrassa	213697	Barcelona	
11	GLDP TORTOSA	410	Dimarts	92400	Supermerca	Tortosa	34432	Tarragona	
12	GLDP VIELHA	290	Dimarts	68000	Petit superi	Vielha	5601	Lleida	
13									

Il·lustració 3 - Estructura Transformada i dades normalitzades

Càrrega → S'ha exportat el fitxer Excel a format CSV per a poder importar-lo des de l'àrea de dades del magatzem de dades. Importació de les dades a taula 'establiment_excel.gldp' des de MySql Workbench amb l'ordre següent:

```
LOAD DATA LOCAL INFILE 'C:/origens/Establiments_export.csv'
INTO TABLE establiment_excel
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
```

Resultat final → Obtenim una taula dimensional amb totes les dades de tots els establiments i amb una clau primària proporcionada per la definició de la taula establiment_excel a la base de dades gldp.

Origen: arxius anuals de productes Productes 20XX.csv

L'origen de dades son arxius de tipus CSV que contenen el catàleg de productes de l'any.

Atès que el camp identificador de producte és clau forana relacionada amb les vendes i hi ha identificadors repetits d'any en any no podem assignar un numero únic nou de producte. La taula productes necessitarà un camp any per poder relacionar el producte del detall_venda amb el producte corresponent de l'any correcte; així doncs la clau principal serà doble l'idproducte i l'any del producte.

Extracció→ Les dades d'origen arriben en arxiu de tipus CSV. Utilitzem tots els camps rebuts, per tant, no apliquem cap filtre de dades però si que hi farem transformacions.

Transformació → A cadascun dels arxius de producte s'ha aplicat:

Normalització de la puntuació de valors numèrics a puntuació americana (substitució de '.' per ',') normalització valor 'Sí' a 'true', inserció de camp d'any del producte amb el valor corresponent a l'any del catàleg. Aquests canvis s'han executat des de l'aplicació Excel.

Càrrega → Un cop obtingut cada arxiu anual CSV amb les dades normalitzades s'ha importat cadascun dels arxius a la taula 'productes_csv.gldp'.

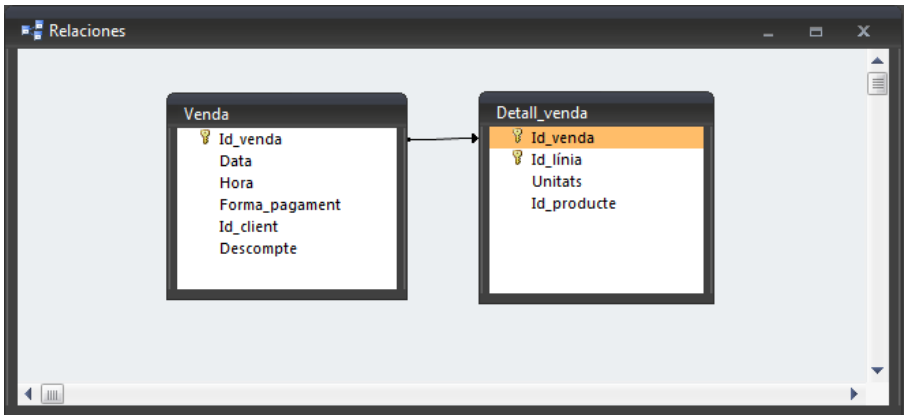
```
LOAD DATA LOCAL INFILE 'C:/origens/Productes2010.csv'
INTO TABLE productes_csv
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'

LOAD DATA LOCAL INFILE 'C:/origens/Productes2011.csv'
INTO TABLE productes_csv
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'

LOAD DATA LOCAL INFILE 'C:/origens/Productes2012.csv'
INTO TABLE productes_csv
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
```

Resultat final → Obtenim la taula 'productes_csv.gldp' amb tots els productes de tots els anys.

Simplificar disseny dimensional → eliminar taula data i assignar camps data, hora i any a taula detall_venda, el camp calculat any facilita la relació amb catàleg de productes.

Origen: arxius Access de dades de venda de cada establiment

Cada establiment de la cadena envia un arxiu amb les seves dades de venda. Aquests arxius son bases de dades tipus MS Access i contenen dues taules relacionades: Venda i Detall_venda.



Il·lustració 4 - Orígens de dades de vendes proporcionat per cada establiment

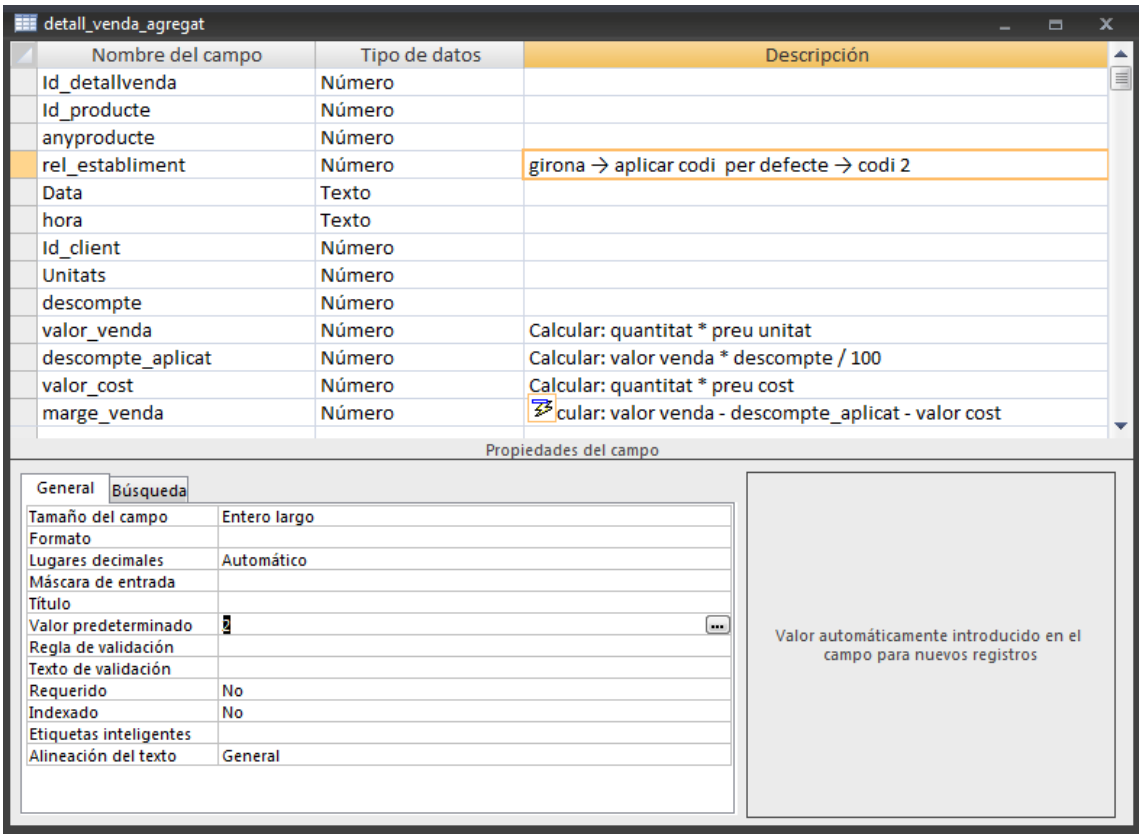
Extracció → Les dades arriben en format base de dades. No usem tots els camps de dades presents en l'origen de dades i segons l'anàlisi realitzats en necessitem d'altres de calculats per a la taula de factual del magatzem de dades. Aplicarem transformacions usant la mateixa aplicació Access per a extreure les dades necessàries detectades en la fase d'anàlisi.

Transformació → Cal aglomerar les dades de la taula venda i la taula detall_venda dels fitxers Access per a conformar-les segons l'especificació de la taula factual detall_venda de la base de dades final del magatzem. També cal afegir l'identificador d'establiment (identificador obtingut en el procés de les dades d'establiment) per a establir la relació detall_venda amb establiment. I finalment afegir els camps calculats en funció del valor del producte.

De les dues taules d'Access hem d'obtenir els camps següents definits a la Taula factual del magatzem de dades
TABLE `detallvenda_csv`

`id_detallvenda`
`id_producte`
`anyproducte`
`rel_establiment`
`data`
`hora`
`id_client`
`unitats`
`valor_venda`
`descompte_aplicat`
`valor_cost`
`marge_venda`

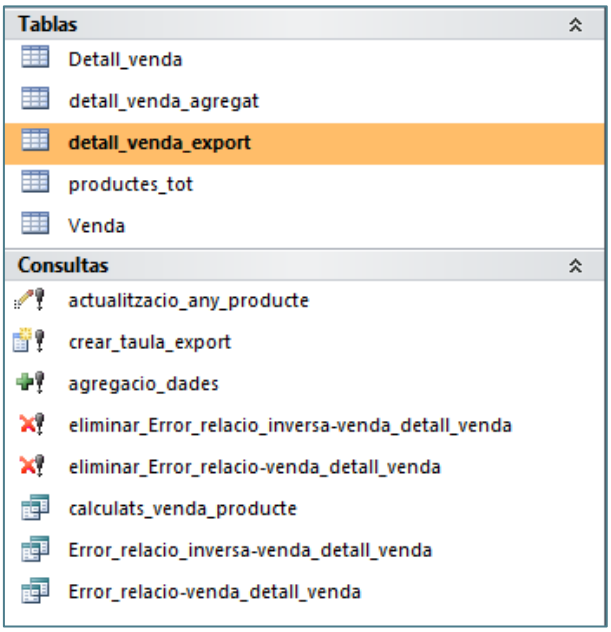
Es crea l'estructura de taula de dades agregades 'detall_venda_agregat' que segueix l'estructura de la taula factual detall venda per tal d'incorporar-hi les dades. Per a cada arxiu origen s'estableix el valor predeterminat d'identificador d'establiment obtingut del procés ETL de l'arxiu d'establiments.



Il·lustració 5 - Estructura de dades de la taula detall_venda_agregat

A part d'obtenir les dades també cal comprovar-ne la integritat i calcular-ne de noves.

Mitjançant consultes d'eliminació, agregació i creació, s'han filtrat, creat i calculat les dades que finalment s'exportaran. A continuació es descriuen els casos.



Il·lustració 6 - Relació de consultes i taules de la base de dades transformada

Integritat de dades de venda: No s'importen dades de venda que no tenen detall de venda o al contrari. S'eliminen els registres no coincidents de la taula bd de treball → aplicar consulta de no coincidents per veure els casos i aplicar consulta d'eliminació de registres no coincidents.

Sentències SQL de les consultes corresponents: eliminar_Error_relació-venda_detall_venda
DELETE Venda.Id_venda FROM Venda WHERE (((Venda.Id_venda) In (SELECT Venda.Id_venda FROM Venda LEFT JOIN Detall_venda ON Venda.[Id_venda] = Detall_venda.[Id_venda] WHERE (((Detall_venda.Id_venda) Is Null)))));
DELETE Detall_venda.Id_venda FROM Detall_venda WHERE (((Detall_venda.Id_venda) In (SELECT Detall_venda.Id_venda FROM Venda RIGHT JOIN Detall_venda ON Venda.[Id_venda] = Detall_venda.[Id_venda] WHERE (((Venda.Id_venda) Is Null)))));

Agregació dels camps de dades necessaris per a la taula factual del magatzem a una sola taula→ consulta d'agregació de dades.

Sentència SQL de la consulta: agregació_dades
INSERT INTO detall_venda_agregat (Id_detallvenda, Id_producte, Data, Hora, Id_client, Unitats, Descompte) SELECT Detall_venda.Id_venda, Detall_venda.Id_producte, Venda.Data, Venda.Hora, Venda.Id_client, Detall_venda.Unitats, Venda.Descompte FROM Detall_venda LEFT JOIN Venda ON Detall_venda.Id_venda = Venda.Id_venda;

Inserció del valor calculat d'any del producte → consulta d'actualització any producte.

Sentència SQL de la consulta: actualització_any_producte
UPDATE detall_venda_agregat SET detall_venda_agregat.anyproducte = Year([detall_venda_agregat]![Data]);

Inserció de valors calculats relacionats amb el valor del producte → importació de taula total de productes i actualització de dades relacionades. Revisió d'integritat de id_producte i anyproducte entre detall venda i taula de productes completa.

Sentència SQL de la consulta: crear_taula_export
SELECT detall_venda_agregat.Id_detallvenda, detall_venda_agregat.Id_producte, detall_venda_agregat.anyproducte, detall_venda_agregat.rel_establiment, detall_venda_agregat.Data, detall_venda_agregat.hora, detall_venda_agregat.Id_client, detall_venda_agregat.Unitats, [detall_venda_agregat]![Unitats]*[productes_tot]![preu_venda] AS valor_venda, [detall_venda_agregat]![descompte]*[productes_tot]![preu_venda]/100 AS descompte_aplicat, [detall_venda_agregat]![Unitats]*[productes_tot]![preu_cost] AS valor_cost, [Valor_venda]-[valor_cost]-[descompte_aplicat] AS marge_venda INTO detall_venda_export FROM detall_venda_agregat INNER JOIN productes_tot ON

```
(detall_venda_agregat.anyproducte = productes_tot.anyproducte) AND
(detall_venda_agregat.Id_producte = productes_tot.idproducte);
```

S'ha programat la sentència per a que obviï els valors de venda que tinguin productes inexistents al catàleg total de productes. VEURE NOTA SOBRE LES DADES DE PRODUCTES DE 2009

Després d'aquestes transformacions s'obté la taula 'detall_venda_export' amb tots els atributs definits en el disseny dimensional completats amb les dades adients.

En aquest punt ja es tenen les dades necessàries per a la càrrega a la base de dades final del magatzem.

Aquesta taula s'exporta en format CSV i es normalitza la puntuació dels valors numèrics a format anglosaxó (substitució de coma decimal per punt decimal)

Càrrega → carreguem les dades agregades a la taula del magatzem de dades amb la sentencia 'LOAD DATA LOCAL INFILE'.

Resultat final → Un cop fet aquest procés per a tots els arxius de vendes obtenim la taula 'detall_venda.gldp' amb totes les dades vàlides de detall venda de tots els establiments.

Llista resum de procés ETL per a tot origen de dades Establiment.accdb

1. **Copiar estructura de taula de dades agregades detall_venda_agregat**
2. **Establir identificador d'establiment per defecte al camp establiment**
3. **Copiar taula de productes tot**
4. **Copiar consultes de comprovació i transformació de dades**
5. **Executar consultes de comprovació i eliminació de relacions errònies entre les taules detall venda i venda**
6. **Executar consulta d'agregació de dades**
7. **Actualitzar any producte**
8. **Executar consulta de creació de dades d'exportació amb els camps calculats**
9. **Exportar taula a format CSV per importar al magatzem de dades**
10. **Normalització puntuació de valors numèrics a puntuació americana (substitució de '.' per ',')**

Aquest procés s'haurà d'aplicar a cada arxiu de base de dades dels establiments del grup.

Notes sobre les fonts de dades

L'enunciat afirma que hi ha dades proporcionades pels establiments que encara que es refereixen al mateix subjecte no s'expressen de la mateixa manera, uns establiments informen valors absoluts mentre que d'altres informen valors de mitjana. A les bases de dades de venda que els establiments ens trameten, no hi ha el valor expressat de manera diferent. I als arxius d'establiments.xls i productesX.CSV tampoc no es veu cap camp susceptible de ser equívoc en aquest sentit. Atès que no hi ha cap heterogeneïtat en les dades obvio aquest punt de l'enunciat.

L'establiment Figueres GLDP ha tramés l'arxiu de dades de venda amb totes les dades incoherents; feta la comprovació d'integritat referencial entre els registres de Venda i els de Detall_venda resulta que no hi ha cap registre corresponent. En d'altres establiments el nombre d'errors d'aquest tipus és manté en un rang de més de 10 a menys de 1000 errors.

En relació a la categorització de clients A/B/C (80/15/5) les dades rebudes dels establiments no contenen en general identificació de client (la major part de registres tenen aquest camp a zero) fer una categorització de clients per aquest atribut no es factible.

S'observen als arxius de vendes valors de nombre d'unitats venudes negatius → Assumeixo que son retorns de mercaderia o errors de caixa que s'han rectificat, per tant, mantinc els valors negatiu d'unitats.

S'observen als arxius de vendes dades dels anys 2009, com que GLDP no proporciona catàleg de productes de l'any 2009 no s'importaran les dades de vendes amb data 2009 atès que no poden ser relacionades adequadament.

Les dades d'habitants de població obtingudes del portal IDESCAT son de data 1 de gener de 2011 atès que la última consulta efectuada al portal en data 1 de gener de 2013 les proporciona actualitzades en aquella data.

III - Errors genèrics al procés ETL

Tipus d'errors previsibles	Actuacions a fer.
Manca d'algun dels orígens de dades	Assegurar que tots els arxius d'extracció necessaris son presents en la fase d'inici del magatzem i també en cada fase d'actualització. Si no es compleix → reclamar l'arxiu necessari a qui correspongui i posposar procés ETL
Error de connexió o elèctric durant la càrrega de dades	La base de dades del magatzem ha de tenir còpia de seguretat. → restaurar la base de dades del magatzem amb la còpia de seguretat.
Manca d'integritat referencial en orígens de dades	Cal comprovar la integritat referencial dels registres relacionats per evitar registres *penjats*. Aplicar el procés ETL programat per evitar aquest error.
Absència de dades. Existeix l'arxiu origen però no conté registres.	Comprovar amb establiment origen si es tracta d'un error de tramesa o es correcte (p.e. període sense dades noves) No aplicar el procés ETL si realment no hi ha dades, o esperar a rebre l'arxiu correcte

Volum de dades gestionades pel magatzem de dades amb aquest procés ETL completat.
Registres totals importats de detall venda : 898.752 **Registres totals importats de productes : 3.089** **Registres totals importats de establiments: 11**

En aquest punt tenim totes les dades que ens interessen carregades a les taules de la base de dades del magatzem de dades. Podem procedir a la realització dels informes sol·licitats.

IV - Informes realitzats

Es mostren il·lustracions dels Informes realitzats

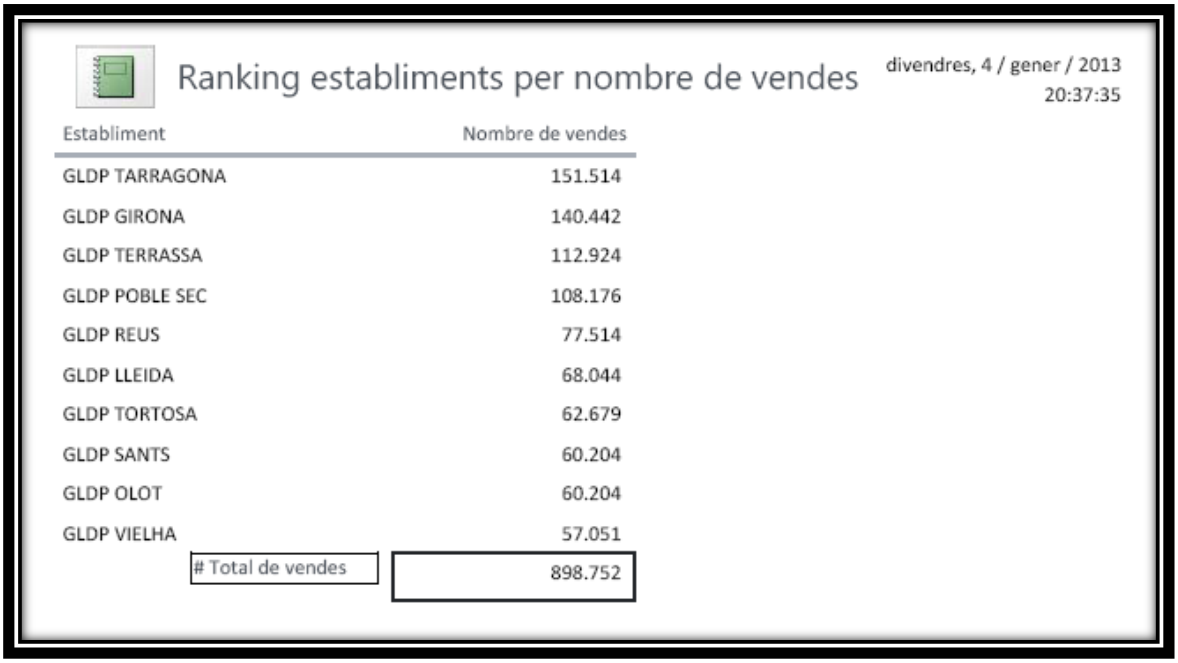
1. Informe de total de vendes i marge net del grup

Informe total de vendes i marge net del Grup GLDP

divendres, 4 / gener / 2013
20:36:36

Suma de valor venda	Suma de marge de venda
4.526.757.157,06 €	1.057.330.792,67 €

2. Ranking d'establiments per vendes i volum total de vendes



Ranking establiments per nombre de vendes

divendres, 4 / gener / 2013
20:37:35

Establiment	Nombre de vendes
GLDP TARRAGONA	151.514
GLDP GIRONA	140.442
GLDP TERRASSA	112.924
GLDP POBLE SEC	108.176
GLDP REUS	77.514
GLDP LLEIDA	68.044
GLDP TORTOSA	62.679
GLDP SANTS	60.204
GLDP OLOT	60.204
GLDP VIELHA	57.051
# Total de vendes	898.752

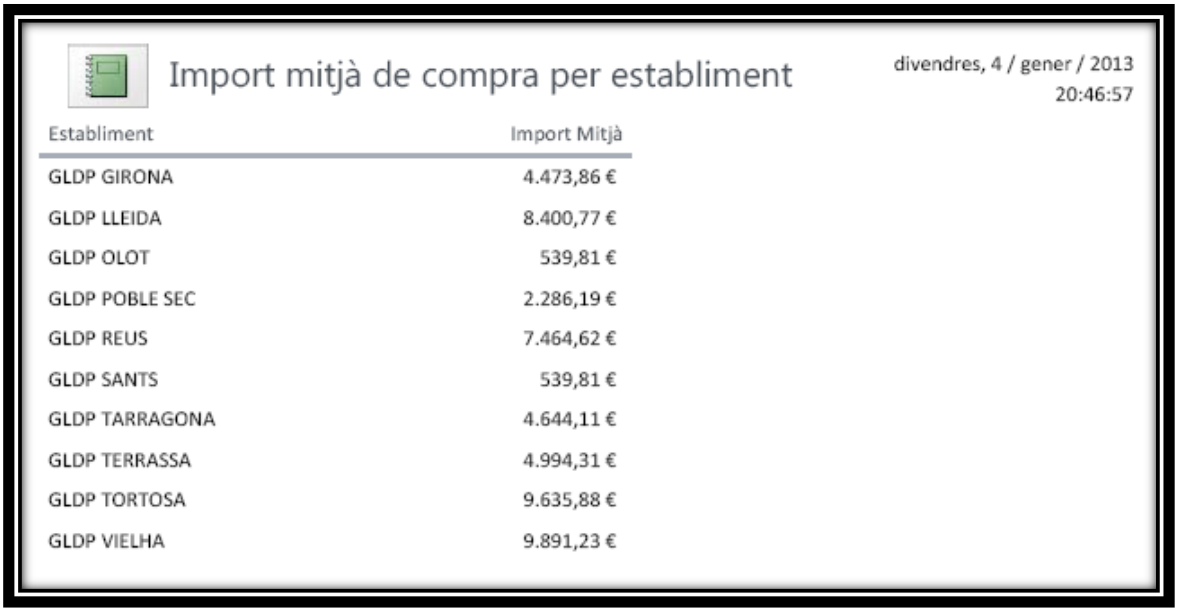
3. Total de vendes per demarcació

Total de valor de vendes per demarcació

divendres, 4 / gener / 2013
21:44:05

demarcacio	SumaDevalor_venda
Barcelona	843.786.251,30 €
Girona	660.816.621,48 €
Lleida	1.135.927.027,80 €
Tarragona	1.886.227.256,48 €
4	

4. Import mitjà de compra per establiment

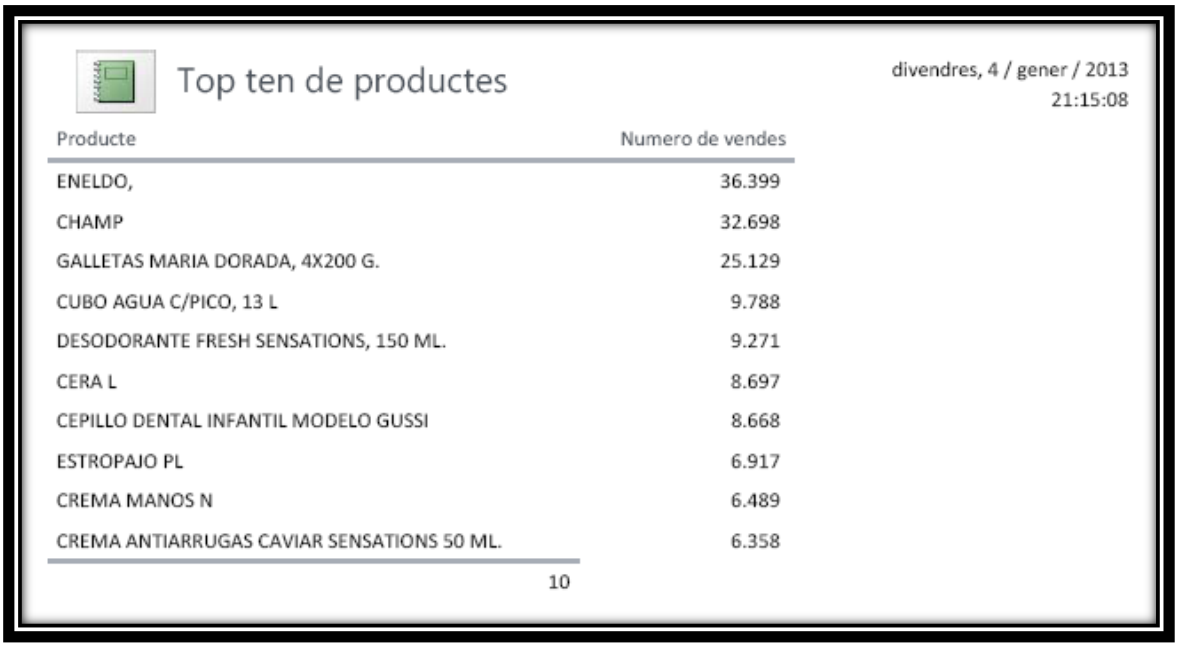


Import mitjà de compra per establiment

divendres, 4 / gener / 2013
20:46:57

Establiment	Import Mitjà
GLDP GIRONA	4.473,86 €
GLDP LLEIDA	8.400,77 €
GLDP OLOT	539,81 €
GLDP POBLE SEC	2.286,19 €
GLDP REUS	7.464,62 €
GLDP SANTS	539,81 €
GLDP TARRAGONA	4.644,11 €
GLDP TERRASSA	4.994,31 €
GLDP TORTOSA	9.635,88 €
GLDP VIELHA	9.891,23 €

5. Top ten de productes



Top ten de productes

divendres, 4 / gener / 2013
21:15:08

Producte	Numero de vendes
ENELDO,	36.399
CHAMP	32.698
GALLETAS MARIA DORADA, 4X200 G.	25.129
CUBO AGUA C/PICO, 13 L	9.788
DESODORANTE FRESH SENSATIONS, 150 ML.	9.271
CERA L	8.697
CEPILLO DENTAL INFANTIL MODELO GUSSI	8.668
ESTROPAJO PL	6.917
CREMA MANOS N	6.489
CREMA ANTIARRUGAS CAVIAR SENSATIONS 50 ML.	6.358

10

V - Conclusions

Sobre l'etapa de planificació

Dificultats

He pogut comprovar que al no haver fet l'assignatura de *Gestió d'organitzacions i projectes informàtics* aquestes primeres fases de determinació d'objectius i organització del projecte m'han suposat una dificultat inesperada, encara que també m'ha aportat l'aprenentatge de nous coneixements.

També he tastat les diferències entre planificació prevista i realització de tasques, sobretot en l'etapa d'anàlisi i de implementació, les fites previstes i el cronograma previst s'han desviat significativament per diverses causes (comentades als apartats corresponents a continuació)

Aprenentatges

En aquesta etapa de planificació he pogut treballar competències de determinació de l'abast i objectius d'un projecte i també l'anàlisi de requeriments; també he practicat l'ús d'eines de planificació i organització, eines de redacció documental i eines de creació de diagrames i retoc d'imatges.

Sobre l'etapa d'anàlisi i disseny

Dificultats

Al començar l'etapa d'anàlisi i disseny em vaig trobar bloquejat, el desconeixement de la matèria que havia de tractar feia que no sabés per on començar i la tasca a fer semblava massa gran. Això va comportar un incompliment de terminis amb la planificació. També vaig trobar que no treballar en equip dificulta la sortida a aquests atzucacs en que un es pot trobar.

Aprenentatges

Des del punt de vista de gestió de projectes he aprés a atacar els problemes d'un en un a abstraure'm d'un objectiu gros aparentment inabastable per desconegut i centrar-me en una tasca concreta que permeti fer un pas i adquirir un coneixement que després et porta al següent pas. "*Tot viatge comença amb un pas"*

Des del punt de vista de coneixements tècnics en aquesta fase del projecte és on he adquirit els conceptes nous relacionats amb la modelització d'un magatzem de dades.

Cal fer notar que el model dimensional del magatzem de dades no segueix la normalització habitual en les bases de dades relacionals. Per exemple, atributs amb valors textuals repetitius com el nom de la població o el proveïdor, segons l'autor R. Kimball, no es converteixen en noves dimensions relacionades amb la taula factual en un magatzem de dades.

Els problemes associats amb l'entrada de dades repetitives errònies que en el model estàndard relacional E-R es solucionen mitjançant la normalització de taules relacionades, en el model dimensional per a magatzems de dades, se solucionen durant al procés de càrrega atès que es revisen i transformen les dades. Aquest fet evita que es creïn entitats/dimensions

innecessàries i simplifica l'esquema general de la base de dades del magatzem i en conseqüència es simplifica també la sintaxi de les consultes que s'hi realitzaran.

També cal assenyalar que el disseny de la base de dades que carregarà les dades finals ha d'estar enfocat a gestionar una sèrie de dades d'un procés de negoci que es vol estudiar, i no a operar amb les dades transaccionals del funcionament d'una corporació.

Sobre l'etapa d'implementació

Dificultats i Aprenentatges

Al treballar la fase ETL he confirmat la problemàtica amb els arxius origen de dades que em plantejava abans del començament del projecte al sol·licitar aquesta àrea de treball de final de carrera. Veig que la fiabilitat i perdurabilitat del model ETL i per tant del correcte funcionament d'un magatzem de dades, és directament proporcional a la fiabilitat dels orígens de dades. Només una sòlida i fiable font d'orígens de dades permet que el model ETL perduri en el temps.

En conseqüència el fet més probable és que per a mantenir operatiu un magatzem de dades calgui una actualització constant del model ETL i una revisió curosa dels orígens de dades.

La tecnologia a usar per explotar la base de dades final i per a executar les tasques OLAP m'ha representat un problema, doncs a la màquina virtual instal·lada l'eina de Pentaho no m'ha funcionat. Finalment vaig optar per seguir amb la solució Microsoft Access, segurament menys potent i menys adaptada a les tasques d'anàlisi i reporting requerides per al projecte, però és la tecnologia a la qual tenia accés real i ha resultat funcional; per això, en el capítol 3 es descriuen mètodes d'exportació i de càrrega de dades per a MySql Workbench encara que finalment he finalitzat el projecte amb la solució de MS Access. La càrrega de dades final s'ha fet a la base de dades magatzem.accdb i he fet els informes amb aquesta aplicació. En la màquina virtual queda creada la base de dades amb Mysql i també tots els arxius de l'àrea de recopilació de dades i la base magatzem.accdb amb els informes i les dades consolidades.

Línies d'evolució futura

Assegurar la fiabilitat de les dades de venda doncs com s'ha vist en l'anàlisi de dades hi ha incoherències i manca d'integritat de dades. Es recomana revisar el sistema de recollida de dades de venda als establiments GLDP.

Revisar la tecnologia de la solució informàtica, migrar a un sistema de base de dades més potent i que faciliti el control d'accés dels usuaris (tipus administrador o consultor) per exemple SQL Server 2012 Data Warehousing i implementar en aquesta solució procediments automatitzats per a la normalització de dades.

Si el client ho considera, seria una millora obtenir la llicència del software de Pentaho Business Intelligence atès que conté eines molt potents i adequades per a l'anàlisi, gestió i explotació de dades.

Conclusions finals

L'experiència d'afrontar un treball de final de carrera ha estat per moments incòmode i difícil però finalment enriquidora atès que he treballat noves competències en planificació de projectes i nous coneixements i experiències.

Vaig sol·licitar aquesta àrea del Treball de Final de Carrera en magatzem de dades per intentar respondre incògnites sobre la gestió eficient de fonts de dades per a obtenir informació significativa des d'un punt de vista corporatiu global i gràcies a aquest projecte he començat a conèixer les eines i procediments del mon del Data Warehousing i el Business Intelligence.

VI - Glossari

Base de dades operacional: Base de dades destinada a gestionar les dades de les transaccions diàries en una organització.

Business Intelligence: Conjunt d'estratègies i eines enfocades a l'administració i creació de coneixement mitjançant l'anàlisi de dades existents a una organització o empresa.

Data Warehousing: veure Magatzem de dades

Dimensió: Punt de vista utilitzat en l'anàlisi d'un cert fet.

ETL (Extract, Transform and Load): És el procés que permet a les organitzacions moure dades des de múltiples fonts, transformant-les, depurant-les i carregant-les en una altra base de dades, data mart o magatzem de dades per a analitzar o bé, en un altre sistema operacional per recolzar un procés de negoci.

Fet: mesura de negoci objecte d'anàlisi usualment numèrica i additiva.

Magatzem de dades: Conjunt de Bases de dades i processos que integren dades de diferents fonts amb informació històrica i que té com a objectiu principal fer de suport en la presa de decisions.

OLAP (On-Line Analytical Processing): Aplicacions que permeten que un usuari accedeixi via navegador d'Internet a veure, manipular i analitzar bases de dades multidimensionals.

Taula dimensional: estructura de dades tipus taula amb atributs descriptius i clau principal.

Taula factual: estructura de dades tipus taula amb dades numèriques sobre un fet i claus foranes a les taules dimensionals.

VII - Bibliografia

The Data Warehouse Toolkit, 2nd Edition (9780471200246) - *Ralph Kimball, Margy Ross*

Documentació de la Universitat Oberta de Catalunya ┐

Bases de Dades I - XP05/05002/00492 - *Jaume Sistac Planas*
Enginyeria del programari - XP03/05060/02078 - *Benet Campderrich Falgueras*
Recerca informàtica, SL
Tècniques de Desenvolupament de Programari - XP02/05049/00099 - Universitat Oberta de Catalunya
Treball de Final de Carrera - XP08/19018/00443 - Gestió i desenvolupament de projectes, Redacció de textos cientificotècnics, Presentació de documents i elaboració de presentacions - Universitat Oberta de Catalunya

Web ┐

Data Warehousing Concepts Oracle9i Data Warehousing Guide

VIII - Annex 1 - Enunciat del TFC amb els requisits del client GLDP

Títol: Construcció i explotació d'un magatzem de dades per a l'anàlisi de vendes d'una cadena de supermercats

Enunciat

Davant la situació de crisi actual i la constant disminució de les vendes, el Grup Líder en Distribució de Proximitat (GLDP) ha decidit fer un estudi a Catalunya per determinar la continuïtat de la seva xarxa d'establiments i conèixer millor el comportament de les vendes per tal de capgirar-ne l'evolució negativa dels darrers anys.

El problema amb el que es troba el grup és que, degut a la gestió distribuïda de les compres de cada establiment i a causa de la consolidació trimestral de les vendes, no té accés a totes les dades que necessita per tal de d'engegar les campanyes de màrqueting oportunes que han de de conduir a un increment de vendes en els establiments i períodes més fluixos.

És per això que, GLDP ha decidit encarregar-nos com a consultoria externa independent, la creació d'un magatzem de dades, mitjançant el qual es pugui obtenir, com a mínim, la següent informació:

• Total vendes i marge net del grup
• Import mitjà de compra per soci i establiment
• % de vendes respecte el total del grup (per establiment i per demarcació)
• Preus màxims i mínims per tipologia d'establiment i característiques del producte.
• Rànking d'establiments per nombre de vendes i volum total
• "Top ten" de productes
• % de vendes de "marques blanques" per habitants
• Categorització de Clients A/B/C
• Anàlisi de compra per impuls
• Distribució setmanal i estacionalitat de vendes

Tota aquesta informació es proporcionarà dintre d'una temporalitat a nivell de mes i any. Es podrà consultar de forma agregada, per demarcació territorial, tipus d'establiment i família de productes.

Addicionalment, haurem de proporcionar un conjunt predefinit d'informes on es mostri la informació sol•licitada i qualsevol altre que pugui ser útil per al GLDP.

GLDP ens proporcionarà tota la informació relativa a la seva activitat recollida en els fitxers següents:

• Catàleg de productes: Un arxiu per cadascun dels anys, amb tots els productes oferts en algun dels seus punts de venda
• Relació d'establiments: Un arxiu amb la fitxa descriptiva de cada establiment.
• Detall de vendes: Un arxiu per cada establiment amb la relació de vendes.

Ens adverteixen que degut a que la informació s'ha extret de diferents sistemes, hi ha tres formats diferents d'arxius (CSV, XLS i MDB).

A més, ens indiquen que, tot i tenir una estructura semblant, no tots els establiments ens donen la mateixa informació, havent establiments que proporcionen dades en valors absoluts i d'altres valors mitjos que caldrà unificar.

També ens demanen que per a que les dades siguin més realistes, el nombre d'habitants (que podem obtenir de l'IDESCAT) el calcularem com a la mitjana de valors a 1 de gener de l'any i l'1 de gener de l'any següent.

Objectius

L'objectiu principal del projecte és adquirir experiència en el disseny, construcció i explotació d'un magatzem de dades a partir de la informació disponible en una base de dades transaccional.

Descripció del treball a realitzar

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

Pla de treball i anàlisi preliminar de requeriments

Al principi del curs es demanarà a l'estudiant un pla de treball on s'indicarà la planificació estimada de les diferents tasques a realitzar per dur a terme el projecte.

L'alumne lliurarà, també, un document d'anàlisi preliminar (no detallat) amb l'enumeració i breu descripció dels elements d'anàlisi identificats (dimensions,atributs, indicadors, etc.) que estaran disponibles per als usuaris i el nombre d'informes aproximat que s'implementaran i contingut dels mateixos. També s'analitzaran les fonts de dades operacionals proporcionades que serviran per carregar cadascun dels elements d'anàlisi.

Anàlisi de requeriments i disseny conceptual i tècnic

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

Implementació

Aquesta fase constarà de les següent tasques:

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

Coneixements previs

• Conceptes generals de bases de dades.
• Conceptes bàsics d'HTML.
• Coneixement del llenguatge PL/SQL.

Requeriments de maquinari i programari

• Sistema operatiu: Windows XP o Linux (recomanat Ubuntu).
• Bases de dades: MySQL
• Programació: PL/SQL de Oracle.
• Recomanat: MySQL WorkBench, ERWin, Suite Pentaho o MS-Visio com eina CASE.

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