Universitat Oberta de Catalunya

uoc.edu

# Sistema d'intel·ligència de negoci per a l'anàlisi publicitari

**Oriol Subirana Perdiguer**

Màster en Enginyeria Informàtica

Business Intelligence

Consultor: **David Amorós Alcaraz**

Responsable de l'assignatura: **María Isabel Guitart Hormigo**

01/2018

**FITXA DEL TREBALL FINAL**

| | |
|---|---|
| **Títol del treball:** | *Sistema d'intel·ligència de negoci per a l'anàlisi publicitari* |
| **Nom de l'autor:** | *Oriol Subirana Perdiguer* |
| **Nom del consultor/a:** | *David Amorós Alcaraz* |
| **Nom del PRA:** | *María Isabel Guitart Hormigo* |
| **Data de lliurament (mm/aaaa):** | *01/2018* |
| **Titulació o programa:** | *Màster en Enginyeria Informàtica* |
| **Àrea del Treball Final:** | *Business Intelligence* |
| **Idioma del treball:** | *Català* |
| **Paraules clau** | *BI, publicitat, Pentaho* |


**Resum del Treball (màxim 250 paraules):**

En els darrers anys la intel·ligència de negoci (*Business Intelligence*) s'ha convertit en un factor estratègic per a les organitzacions, generant un potencial avantatge competitiu, que no és un altre que proporcionar informació privilegiada per respondre als problemes del negoci.

El primer objectiu del projecte es trobar una solució *Open Source* de *Business Intelligence* que possibiliti l'anàlisi de la informació generada per centres comercials situats en diferents zones geogràfiques i donar resposta a les preguntes plantejades.

*Pentaho* ha sigut l'eina BI escollida i ens ha permès explotar les dades a partir del *Data Warehouse* creat mitjançant processos *ETL* i cubs *OLAP*.

Els objectius del projecte s'han assolit amb èxit ja que s'han pogut respondre totes les qüestions plantejades.

**Abstract (in English, 250 words or less):**

In the past few years, *Business Intelligence* has become a strategic factor for the companies, generating a potential competitive advantage, which is none other than providing insider information to respond to business problems.

The first objective of the project is to find an *Open Source Business Intelligence* solution that will allow the analysis of the information generated by shopping centres located in different geographical areas and respond to the questions posed.

*Pentaho* has been the chosen BI tool and has allowed us to exploit the data from the *Data Warehouse* created through *ETL* processes and *OLAP* cubes.

The objectives of the project have been successfully achieved since all the questions raised have been answered correctly.

# Índex

# Llista de figures

# 1. Introducció

## 1.1. Context i justificació del Treball

La finalitat del Treball Final de Màster presentat en aquesta memòria és crear un entorn Business Intelligence (BI) que possibiliti l'anàlisi de la informació generada durant una campanya publicitària basada en la producció de petits anuncis en forma d'impactes visuals en pantalles o monitors situats dins varis centres comercials.

El marxandatge visual ha demostrat ser un incrementador en clients i en l'efecte sobre les vendes. Les seves radicals funcions distintives ajuden al disseny de tendes de tot tipus de comerç a nivell global on, els resultats demostren, que el disseny d'un establiment tant intern com a extern també sorgeix de la part seductora cap als clients. Els seus elements cobren acció en les decisions dels compradors posat que de manera inconscient s'interessen per al facilitat de comprar.

Les noves tecnologies han fet evolucionar el concepte de marxandatge visual incorporant noves formes de captar l'atenció des de la reproducció d'anuncis en tot tipus de pantalles interconnectades, publicitat dins les apps o en les xarxes socials més utilitzades.

Tot i que els avantatges immediats d'utilitzar aquestes tecnologies són molt clars, transformar les dades que generen en informació pel negoci, segueix sent difícil i desconegut per la majoria de companyies.

I és en aquest punt, on es fa palesa la necessitat de crear un entorn Business Intelligence que possibiliti l'anàlisi de tota aquestes dades per a convertir-la en informació útil per al desenvolupament del negoci.

## 1.2. Objectius del Treball

L'objectiu d'aquest treball és el disseny i implementació d'un sistema de Business Intelligence que faciliti l'adquisició, l'emmagatzemament i l'explotació de dades obtingudes de la campanya publicitària portada a terme en ubicacions diferents.

Els objectius específics del treball són:

1. Dissenyar un magatzem de dades (Data Warehouse) que permeti emmagatzemar la informació adquirida dels diferents orígens.

2. Implementar aquest magatzem de dades i programar els processos ETL (extracció, transformació i càrrega) que permetin alimentar el Data Warehouse a partir dels fitxers base facilitats.

3. Analitzar les diferents plataformes BI Open Source disponibles al mercat que ens permetin explotar la informació emmagatzemada.

4. Triar i implantar una d'aquestes eines Open Source de tal forma que es disposi d'una capa de programari per l'anàlisi de la informació.

## 1.3. Enfocament i mètode seguit

Per assolir els objectius s'analitzaran les diferents plataformes BI Open Source disponibles al mercat per trobar la que s'ajusti més a la fi del projecte. Després d'una primera cerca la conclusió és que les plataformes BI Open source més utilitzades són: BIRT, JasperReport, Pentaho i SpagoBI.

La decisió d'escollir una de les plataformes, es basarà en les diferents eines i solucions que aporti cada una d'elles. Principalment es tindrà en compte la compatibilitat de cada plataforma amb els orígens de les dades, la dificultat

d'implementació que comporti cada una d'elles i la comunitat que hi hagi darrera de cada una de les plataformes.

Un cop escollida la plataforma que millor s'ajusti a les necessitats del projecte, es realitzarà la mateixa feina per el SGDB que funcioni de magatzem (Data Warehouse) de les dades.

## 1.4 Planificació del Treball

Els recursos necessaris per a la realització del projecte es divideixen en tres grans grups: Recursos Humans, Recursos de Software i Recursos d'Entorn.

Al tractar-se d'un treball personal, els recursos humans es basen en un sol autor del projecte amb la col·laboració necessària del consultor de l'assignatura.

Pel que fa als Recursos de Software s'utilitzarà una plataforma BI Open Source per a la gestió i transformació de les dades en informació i un sistema de gestió de bases de dades relacional i lliure per dur a terme el disseny del magatzem de dades (Data Warehouse).

Finalment el Recurs d'Entorn es comprendrà d'un ordinador portàtil amb sistema operatiu Mac Os X amb connexió a internet.

Un cop establerts els recursos del projecte, es confecciona una planificació temporal de les tasques a utilitzar. Les tasques es defineixen en funció de les fites parcials establertes en cada una de les PAC.

La següent taula mostra la planificació del projecte:

| | Tasques i activitats | Durada | Data inici | Data fi | Fites |
|---|---|---|---|---|---|
| | **1. Pla de Treball** | **12 dies** | **21/09/2017** | **02/10/2017** | |
| 1.1 | Anàlisi i comprenció de l'enunciat | 1 | 21/09/2017 | 21/09/2017 | |
| 1.2 | Anàlisi de la bibliografia | 3 | 22/09/2017 | 24/09/2017 | |
| 1.3 | Cerca de les eines BI Open Source | 4 | 25/09/2017 | 28/09/2017 | |
| 1.4 | Cerca per a la realització del Pla de Treball | 1 | 29/09/2017 | 29/09/2017 | |
| 1.5 | Redacció de la part proporcional a la memòria del projecte | 3 | 30/09/2017 | 02/10/2017 | PAC 1 |
| | **2. Anàlisi de les diferents plataformes BI Open Source** | **11 dies** | **04/10/2017** | **14/10/2017** | |
| 2.1 | Comparació i elecció entre les diferents eines BI Open Source | 5 | 04/10/2017 | 08/10/2017 | |
| 2.2 | Comparació enre els diferents SGBD compatibles amb l'eina BI escollida | 3 | 09/10/2017 | 11/10/2017 | |
| 2.3 | Redacció de la part proporcional a la memòria del projecte | 3 | 12/10/2017 | 14/10/2017 | |
| | **3. Disseny i implementació del Data Warehouse** | **19 dies** | **15/10/2017** | **03/11/2017** | |
| 3.1 | Anàlisi de les dades inicials i les seves interrelacions | 2 | 15/10/2017 | 16/10/2017 | |
| 3.2 | Definició de la base de dades: camps, taules i connexions | 2 | 17/10/2017 | 18/10/2017 | |
| 3.3 | Instal·lació del SGBD escollit | 2 | 19/10/2017 | 20/10/2017 | |
| 3.4 | Disseny del model conceptual del Data Warehouse | 4 | 21/10/2017 | 24/10/2017 | |
| 3.5 | Redacció de la part proporcional a la memòria del projecte | 4 | 25/10/2017 | 28/10/2017 | |
| 3.6 | Implementació del disseny conceptual del Data Warehouse | 5 | 29/11/2017 | 03/11/2017 | |
| | **4. Programar processos ETL** | **24 dies** | **04/11/2017** | **27/11/2017** | |
| 4.1 | Aprenentatge de l'eina escollida | 5 | 04/11/2017 | 08/11/2017 | PAC 2 |
| 4.2 | Comprovació de les dades i proves amb SGBD | 8 | 09/11/2017 | 16/11/2017 | |
| 4.3 | Càrrega de les dades al Data Warehouse | 8 | 17/11/2017 | 24/11/2017 | |
| 4.4 | Redacció de la part proporcional a la memòria del projecte | 3 | 25/11/2017 | 27/11/2017 | |
| | **5. Implementació de BI** | **24 dies** | **28/11/2017** | **21/12/2017** | |
| 5.1 | Anàlisis tècniques BI | 2 | 28/11/2017 | 29/12/2017 | |
| 5.2 | Anàlisi per a la creació dels cubs OLAP | 2 | 30/11/2017 | 01/12/2017 | |
| 5.3 | Creació dels cubs OLAP | 3 | 02/12/2017 | 04/12/2017 | PAC 3 |
| 5.4 | Anàlisi tècnic de la plataforma BI escollida | 2 | 05/12/2017 | 06/12/2017 | |
| 5.5 | Anàlisi i extracció dels indicadors clau | 4 | 07/12/2017 | 10/12/2017 | |
| 5.6 | Explotació de les dades mitjançant la plataforma BI | 3 | 11/12/2017 | 13/12/2017 | |
| 5.7 | Disseny i implementació del cuadre de comandament amb la plataforma | 3 | 14/12/2017 | 16/12/2017 | |
| 5.8 | Disseny i execució de bateria de proves | 3 | 17/12/2017 | 19/12/2017 | |
| 5.9 | Correcció / implementació de millores | 1 | 20/12/2017 | 20/12/2017 | |
| 5.10 | Redacció de la part proporcional a la memòria del projecte | 1 | 21/12/2017 | 21/12/2017 | |
| | **6. Conclusions** | **6 dies** | **21/12/2017** | **28/12/2017** | |
| 6.1 | Resposta a les qüestions plantejades a l'enunciat del projecte | 2 | 21/12/2017 | 22/12/2017 | |
| 6.2 | Extreure les conclusions del projecte | 2 | 23/12/2017 | 24/12/2017 | |
| 6.3 | Redacció de la part proporcional a la memòria del projecte | 2 | 27/12/2017 | 28/12/2017 | |
| | **7. Maquetació de la memòria i presentació** | **8 dies** | **29/12/2017** | **07/01/2018** | |
| 7.1 | Maquetació i finalització de la memòria | 3 | 29/12/2017 | 02/01/2018 | |
| 7.2 | Document de presentació | 2 | 03/01/2018 | 04/01/2018 | |
| 7.3 | Preparació de la presentació | 3 | 05/01/2018 | 07/01/2018 | ENTREGA |

## 1.5 Breu sumari de productes obtinguts

Els productes obtinguts amb la realització d'aquest projecte són un *Data Warehouse* implementat en *MySQL* i una estructura de dades en forma de cubs *OLAP* que connecta amb el *Data Warehouse* per fer les consultres i extraccions de les dades desitjades des de qualsevol plataforma de *Business Intelligence* que permeti explotar cubs *OLAP.* En aquest cas, s'ha fet servir *Pentaho CE* i un dels seus *plugins*, anomenat *Saiku Analytics*, per explotar les dades i generar

els informes. Els productes finals resultants entregats juntament amb aquesta memòria són:

- Transformacions ***carga_points*** i ***carga_visits*** que permeten el bolcat de la informació dels excels a la base de dades des de *Pentaho Data Integration*.
- Script de creació de la base de dades ***StagingArea***, disponible a l'annex 1 d'aquesta memòria.
- Script ***loadStagingArea.sql*** que permet la càrrega d'informació a la base de dades ***StagingArea.***
- Script de creació i càrrega a la base de dades **PubliDW**, disponible a l'apartat 3.7 d'aquesta memòria.
- Fitxer **CubPubliDW.xml** amb l'estructura dels cubs OLAP, disponible en l'annex
- Informes generats mitjançant *Saiku Analytics*, inclosos en el capítol 4 d'aquesta memòria

## 1.6 Breu descripció dels altres capítols de la memòria

**Capítol 2. Anàlisi de les eines *BI* i *SGBD*.**
Aquest capítol conté un anàlisi breu sobre les alternatives BI *Open Source* més conegudes dins del mercat de *Business Intelligence* i els SGBD més populars i que millor s'adapten a aquest projecte. En aquest mateix projecte s'explica quines són les eines escollides.

**Capítol 3. Disseny conceptual del *Data Warehouse.***
En aquest capítol, es mostren els passos seguits per dissenyar i implementar el *Data Warehouse* que emmagatzemarà totes les dades inicials. El capítol inclou des de la definició teòrica dels termes utilitzats fins a la transformació *ETL* i la creació i importació de dades al *Data Warehouse*.

**Capítol 4. Explotació de dades.**

En aquest capítol es defineix el procediment per a realitzar l'explotació de dades que facilitarà l'assoliment dels objectius del TFM. Es mostra com definir els cubs *OLAP* amb *Schema Workbench* i com publicar els esquemes per poder accedir des de *Saiku Analytics*. Per últim, aquest capítol també inclou els informes resultants de l'extracció de dades i les respostes a les qüestions plantejades en els objectius del projecte.

# 2. Anàlisi de les eines BI i SGBD

Abans d'escollir quina eina BI Open Source utilitzar per dur a terme el projecte, es realitzarà un petit estudi i comparació de les eines més utilitzades per poder decidir amb fonaments quina d'aquestes eines s'ajusta més a les necessitats finals del projecte.

De la mateixa manera, s'estudiaran les diferents opcions de Sistema de Gestió de Bases de Dades que ens permetin importar les dades d'origen i crear el Data Warehouse. La decisió d'escollir un SGBD o un altre dependrà també de la seva compatibilitat amb l'eina BI escollida.

## 2.1. Comparació d'eines *Business Intelligence Open Source*

L'elecció de l'eina o plataforma *Business Intelligence* que s'utilitzi per a la implementació del *Data Warehouse* és una de les claus principals durant la fase d'anàlisis previ de requeriments i eines de treball. Aquesta eina és la que ens acompanyarà durant tot el projecte i la que ens permetrà implementar el model conceptual que es definirà més endavant. També ens proporcionarà eines per realitzar la càrrega, transformació i extracció de dades mitjançant procediments ETL i finalment ens permetrà extreure informació de les dades utilitzant informes que ens resoldran, o aquest és l'objectiu, les preguntes plantejades a l'enunciat del projecte.

Actualment al mercat hi ha molts productes orientats a la intel·ligència de negoci que ens permeten realitzar les tasques descrites. Per a realitzar aquest petit estudi o comparació ens hem centrat en les eines BI més utilitzades que ofereixin en el seu catàleg solucions integrades (que disposi de tots els mòduls per implementar cada una de les fases del projecte) i que siguin de programari lliure (*Open Source*). També es tindrà molt en compte la facilitat d'aprenentatge i la comunitat d'usuaris de cada una d'elles.

Per els motius anteriors s'ha decidit, coincidint amb la proposta del professorat, que les principals alternatives són:

- ***Pentaho:*** És l'actual líder per excel·lència en quant a solucions de *Business Intelligence Open Source*. Aquesta plataforma combina l'anàlisi de negoci amb la integració de dades que permet als usuaris empresarials a prendre decisions basades en la informació, ciència de dades per crear models de dades robustos i administradors IT per oferir una plataforma segura i escalable per a un ampli conjunt de dades. Pentaho ens proporciona potents eines d'explotació i visualització de dades para la gestió de processos ETL (extracció, transformació i càrrega de dades) mitjançant l'eina *Data Integration*. També permet crear informes interactius, anàlisis multidimensionals de informació (OLAP), mineria de dades i crear quadres de comandament avançats gràcies a l'eina *Dashboard Designer* o el *Community Dashboard Framework.* Tots aquests serveis estan integrats en una plataforma web a la que el usuari pot accedir de manera fàcil i intuïtiva.

- ***SpagoBI:*** Aquesta és l'única plataforma *Business Intelligence* 100% *Open Source*, és a dir, completament gratuïta. L'eina cobreix i satisfà tots els requisits de BI, tant en termes d'anàlisis i de gestió de dades, administració i seguretat. Ofereix solucions per a la presentació d'informes, anàlisis multidimensional (OLAP), mineria de dades, quadres de comandament i consultes ad-hoc. Afegeix també mòduls propis per a la gestió de processos de col·laboració i anàlisis de geo-referència. *SpagoBI* és una plataforma d'integració, i no de producte, ja que no es construeix entorn un conjunt definit d'eines. Té una estructura modular en la qual tots els mòduls es relacionen amb el nucli del sistema, el que garanteix l'harmonia de la plataforma junt a la seva capacitat evolutiva.

- ***Jasper Reports:*** *JasperSoft Business Intelligence Suite*, és la denominació del conjunt d'eines que permeten a les organitzacions generar informació basada en les seves pròpies dades de gestió per a l'avaluació i presa diària de decisions, de forma dinàmica i online. El

*framework* de treball de *JasperSoft* permet integrar fàcilment les diverses fonts de dades disponibles a l'empresa, i mitjançant tècniques d'anàlisis multidimensional obtenir indicadors que, presentats en quadres de comandament i informes dinàmics, proveeixen d'aquesta sensible informació a l'alta gerència. La *Suite* de *JasperSoft* disposa de: *Jasper Reports*, *Jasper ETL, JasperAnalysis, i JasperServer* que és una aplicació Java que actua com a mòdul principal i que proporciona capacitats avançades d'elaboració d'informes, informes ad-hoc, quadres de comandament i administració de permisos, entre altres funcionalitats.

- ***BIRT:*** Aquesta tecnologia disposa de dos components principals: un dissenyador d'informes visuals dins d'Eclipse IDE per a crear informes i anàlisis multidimensionals (OLAP), i un component de rutina per generar informes que poden ser posats en ús en qualsevol entorn Java. El projecte BIRT també inclou un motor de gràfics que està integrat en el dissenyador d'informes i, a més a més, pot ser utilitzat per separat per influir gràfics en una aplicació. Els dissenys d'informes BIRT es desenvolupen en format XML i poden accedir a un cert nombre de fonts de dades diferents, incloent SQL, JDO, XML, Serveis Web, entre altres. El projecte rep suport des d'una activa comunitat d'usuaris a BIRT Exchange i desenvolupadors de Eclipse.org.

Una vegada analitzades les diferents alternatives per desenvolupar el projecte, s'ha decidit que l'eina més adequada és ***Pentaho BI Suite Community Edition.*** Els criteris claus que donen suport a aquesta decisió són els següents:

- Comunitat extensa i participativa d'usuaris.
- Disposa de tots els mòduls per poder dur a terme un projecte BI.
- Consta d'un potent motor ETL simple d'utilitzar per extreure i transformar les dades de diferents fonts de dades.
- Líder mundial en software BI *Open Source.*
- Compatible amb la majoria de sistemes SGBD.

*Pentaho* és una suite molt completa que ens permet poder dur a terme fàcilment cada una de les fases d'un projecte BI mitjançant als diferents mòduls dels que disposa. Concretament, les eines principals que ens ofereix la plataforma escollida són:

- ***Pentaho Data Integration (PDI):*** Ens permet transformar i integrar dades entre sistemes d'informació existents i els *Data Warehouse* que compondran el sistema BI. Aquest mòdul està basat en dos tipus d'objectes: Transformacions ETL i treballs (col·lecció de transformacions).

- ***Pentaho Reporting (PRD):*** Obté i mostra informes dels indicadors de l'organització. Poden ser confeccionats per un usuari final, o estar predefinits per una consulta directa. Aquests informes poden ser visualitzats en formats estàndard com html, pdf o excel.

- ***Pentaho Analysis:*** Permet consultar, explorar i analitzar la informació de l'organització de manera interactiva, podent seleccionar diferents perspectives d'aquesta informació en base a criteris predefinits.

## 2.2. Comparació de Sistemes de gestió de Bases de Dades

El SGBD és necessari per poder crear una base de dades, preferentment relacional, on poder importar les dades originals facilitades en format Excel. Aquesta base de dades és el que anomenem *Data Warehouse* i que construirem mitjançant l'eina *Pentaho Data Integration* amb l'ajuda dels processos ETL.

Per tal de poder escollir el SGBD que millor s'adapti a les necessitats del projecte s'han analitzat els dos sistemes Open Source per excel·lència: *MySQL y PostgreSQL.*

*PostgreSQL i MySQL* són dos dels SGBD Open Source relacionals més utilitzats del mercat. Han competit sempre molt favorablement amb sistemes de bases de dades comercials. *MySQL* es percep com un sistema molt més ràpid però ofereix menys funcions. En canvi, *PostgreSQL* té un conjunt de funcions més extensa. *MySQL* té un punt fort en la seva rapidesa i capacitat per crear nous projectes ràpidament on la principal preocupació és l'optimització de consultes senzilles. Per altre banda, *PostgreSQL* s'ha enfocat tradicionalment en la fiabilitat, integritat de dades i característiques integrades enfocades al desenvolupador.

Els dos SGBD esmentats són compatibles amb *Pentaho Data Integration* i vàlids per al objectiu del projecte. Finalment s'ha optat per utilitzar *MySQL* degut a l'experiència personal del autor del projecte i la familiaritat amb el administració gràfic *MySQL Workbench* que ens permet entre altres coses, gestionar la base de dades de MySQL i utilitzar eines molt útils per al projecte com l'enginyeria inversa que ens permetrà obtenir diagrames de la base de dades.

# 3. Disseny conceptual del *Data Warehouse*

A continuació és realitzarà una breu introducció als conceptes de Data Warehouse i processos ETL.

## 3.1. Definició de *Data Warehouse*

Un Data Warehouse és una base de dades corporativa que es caracteritza per integrar i depurar informació d'una o més fonts diferents, per seguidament processar-la permetent el seu anàlisis des d'infinitat de perspectives i amb grans velocitats de resposta. La creació d'un *Data Warehouse* representa en la majoria d'ocasions el primer pas, des d'un punt de vista tècnic, per implementar una solució completa i fiable de *Business Intelligence.*

L'avantatge principal d'aquest tipus de bases de dades radica en les estructures en les que s'emmagatzema la informació (models de taules en estrella, en floc de neu, entre d'altres). Aquest tipus de persistència de la informació és homogènia i fiable, i permet la consulta i tractament jerarquitzat de la mateixa.

El terme *Data Warehouse* es tradueix literalment com *magatzem de dades* i va ser utilitzat per primera vegada per *Bill Inmon.* Segons va definir el propi *Bill Inmon,* un *Data Warehouse* es caracteritza per ser:

- **Integrat:** les dades emmagatzemades en el *Data Warehouse* s'han d'integrar en una estructura consistent, per el que les inconsistències existents entre els diversos sistemes s'han d'eliminar. La informació es sol estructurar en diferents nivells de detall per adaptar-se a les diferents necessitats dels usuaris.

- **Temàtic:** només les dades necessàries per el procés de generació de coneixement del negoci s'integren des del entorn operacional. Les dades

s'organitzen per temes per facilitat el seu accés i comprensió per part dels usuaris finals.

- **Històric:** el magatzem d'informació d'un *Data Warehouse* existeix per ser llegit, però no modificat. La informació és per tant permanent, significant l'actualització del *Data Warehouse* la incorporació dels últims valors que prendran les diferents variables contingudes en ell sense cap tipus d'acció sobre les dades que ja existien.

Una altre característica dels *Data Warehouse* és que conté metadades, és a dir, dades sobre les dades. Les metadades permeten saber la procedència de la informació, la seva periodicitat d refresc, la seva fiabilitat, entre altre coses.

Des del punt de vista de la intel·ligència de negoci les principals aportacions dels *Data Warehouse* són:

- Proporcionar una eina per a la presa de decisions en qualsevol àrea funcional, basant-se en informació integrada i global del negoci.

- Facilitar l'aplicació de tècniques estadístiques d'anàlisis i modelització per trobar relacions ocultes entre les dades del magatzem; obtenint un valor afegit per el negoci d'aquesta informació.

- Proporcionar la capacitat d'aprendre de dades del passat i de precedir situacions futures en diversos escenaris.

- Simplificar dins del negoci la implementació de sistemes de gestió integral de la relació amb el client.

- Suposa una optimització tecnològica i econòmica en entorns de Centre de Informació, estadística o de generació d'informes amb retorns de la inversió espectaculars.

## 3.2. Processos ETL

Els processos ETL són una part de la integració de dades, però és un element important amb la funció de completar el resultat de tot el desenvolupament de la cohesió d'aplicacions i sistemes.

La paraula ETL correspon a les sigles en angles:

- Extreure (*Extract*)
- Transformar (*Transform*)
- Carregar (*Load*)

El procés ETL consta precisament d'aquestes tres fases: extracció, transformació i carga.

Durant l'etapa d'**extracció**, cobreix l'extracció de dades del sistema d'origen i les fa accessible per a un posterior processament. L'objectiu principal del pas d'extracció és recuperar totes les dades requerides del sistema d'origen amb els recursos mínims possibles. Aquest pas s'ha de dissenyar de manera que no afecti negativament al sistema d'origen en termes de: rendiment, temps de resposta o qualsevol tipus de bloqueig.

El pas de **transformació** aplica un conjunt de regles per transformar les dades de l'origen al destí. Això inclou convertir totes les dades mesurades a la mateixa dimensió utilitzant les mateixes unitats perquè puguin unir-se posteriorment. El pas de transformació també requereix unir dades de diverses fonts, generar agregats, generar claus alternatives, classificar, obtenir nous valors calculats i aplicar regles de validació avançades.

Durant l'etapa de **càrrega**, cal assegurar-se que la càrrega es realitzi correctament i amb els menys recursos possibles. L'objectiu del procés de càrrega és sovint una base de dades. Perquè el procés de càrrega sigui eficaç, és útil desactivar les restriccions i els índexs abans de la càrrega i tornar-los a

habilitar només després de completar la càrrega. El procés ETL necessita mantenir la integritat referencial per garantir la coherència.

La il·lustració següent mostra les fases d'un procés ETL des de les fonts inicials de dades fins als processos analítics d'aquestes al final del procés, que en definitiva és l'objectiu final de les tècniques de *Business Intelligence* que s'estan tractant en aquest projecte.

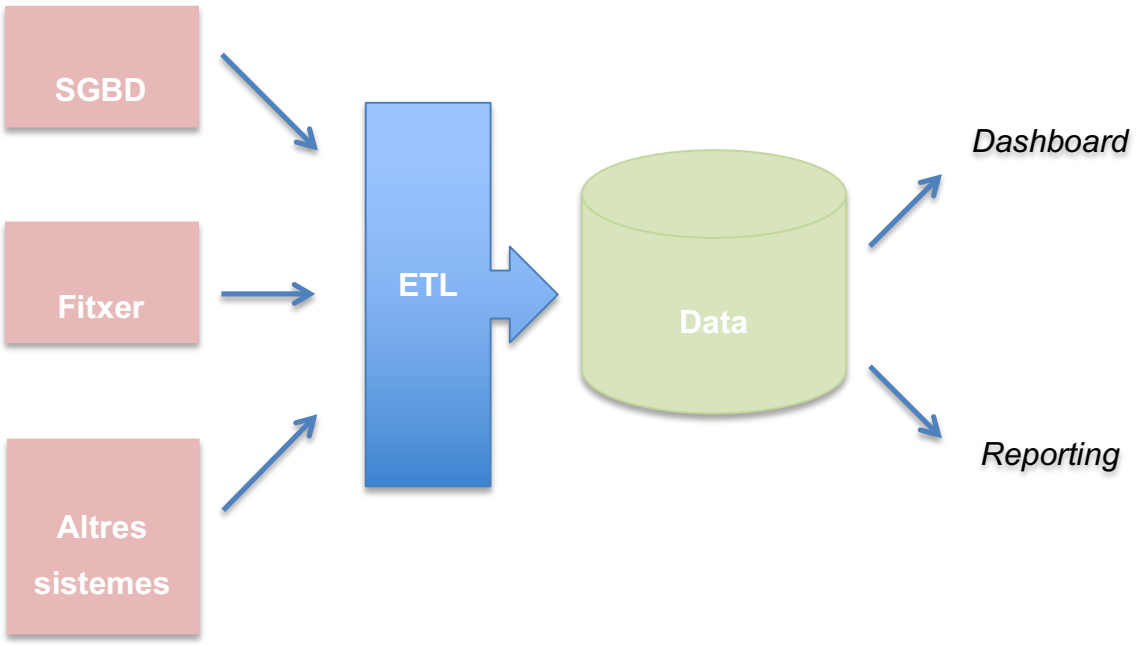



**Fig 1.** Processos ETL

## 3.3. Anàlisis de les dades originals i les seves interrelacions

El primer pas abans de començar amb el disseny i creació del *Data Warehouse* és analitzar les dades originals que es proporcionen al inici del projecte. En aquest cas, es parteix de dos fitxers en format Excel amb la informació de impactes, vendes i visites dels diferents punts de venda.

El primer fitxer facilitat és el ***DATACAMPAIGNIMPACT.xlsx,*** que conté les següents fulles:

- ***Points:*** Zona, ciutat, nom del centre, tipus d'establiment i punt d'impacte.

- ***Products:*** Família i nom dels productes.
- ***Impacts Gran Jonquera SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial Gran Jonquera.
- ***Sales Gran Jonquera SC:*** Ventes per data i producte del centre comercial Gran Jonquera.
- ***Impacts Girocentre SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial Girocentre.
- ***Sales Girocentre SC:*** Ventes per data i producte del centre comercial Girocentre.
- ***Impacts La Maquinista SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial La Maquinista.
- ***Sales La Maquinista SC:*** Ventes per data i producte del centre comercial La Maquinista.
- ***Impacts Mercat del Pla SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial Mercat del Pla.
- ***Sales Mercat del Pla SC:*** Ventes per data i producte del centre comercial Mercat del Pla.
- ***Impacts Les Gabarres SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial Les Gabarres.
- ***Sales Les Gabarres SC:*** Ventes per data i producte del centre comercial Les Gabarres.
- ***Impacts La Fira SC:*** Impactes classificats per dates i punts d'impacte de tots els productes del centre comercial La Fira.
- ***Sales La Fira SC:*** Ventes per data i producte del centre comercial La Fira.

El segon fitxer ***DATACAMPAIGNIMPACT.xlsx***, que conté les fulles següents:

- ***Visits:*** Data i número de visitants per cada un dels centres comercials.

### 3.3.1. Importació de les dades originals a la base de dades MySQL

Primerament es fa una extracció i importació de les dades dels Excels a la base de dades. Per importar les dades a *MySQL* s'utilitzarà l'eina *Pentaho Data Integration* que permetrà crear les taules de forma automàtica a la base de dades i importar la informació provinent dels dos excels. Aquestes primeres taules no tindran cap tipus de restricció de format, ni relació entre elles.

Aquesta primera importació ens permetrà tenir les dades dins de la nostra base de dades i en conseqüència ens facilitarà el procediment de disseny i implementació del *Data Warehouse.*

Les passos a dur a terme en aquesta primera importació són:

1. Creació de la base de dades des de l'eina *MySQL Workbench*. En aquest cas es dirà *tfm_oriol* i l'esquema es dirà *analisi_publicitari.*

2. Des de l'eina *Pentaho Data Integration* crearem la connexió a la nostre base de dades i comprovem que tingui connexió.

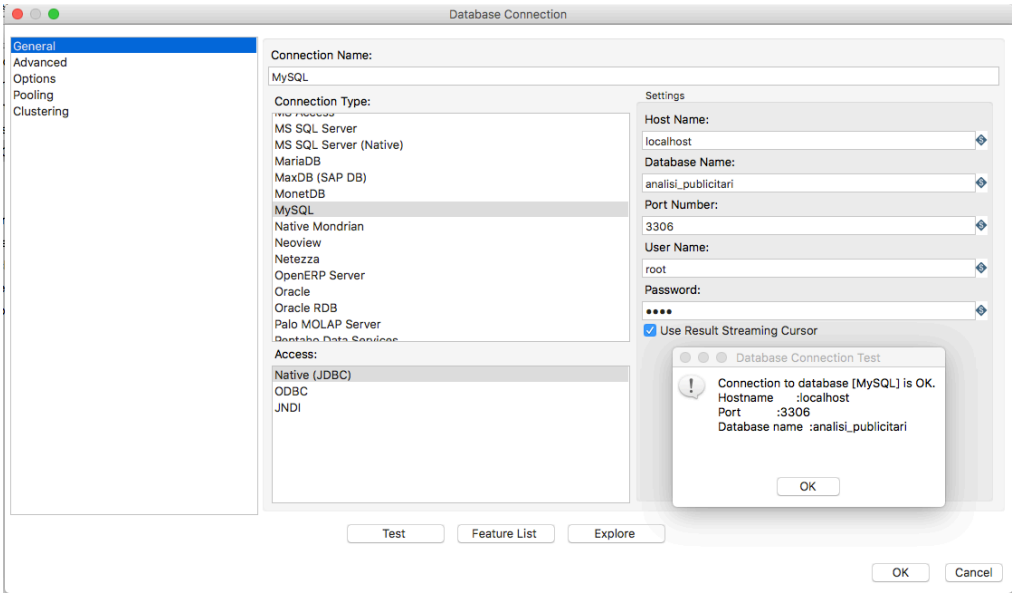



**Fig 2.** Connexió *Pentaho* amb MySQL

3. Creació d'una importació per cada una de les fulles dels excels. Com que el programa ens deixa seleccionar de quina fulla del Excel volem importar les dades, no tindrem problemes en realitzar la importació sense haver de modificar cap dels dos fitxers.


**Fig 3.** Carga inicial *Data Integration*

4. Un cop carregats a la base de dades totes les fulles tindrem totes les taules creades i amb la informació a la BD. Des del administrador de base de dades MySQL Workbench podrem utilitzar la enginyeria inversa per crear el diagrama de la base de dades.

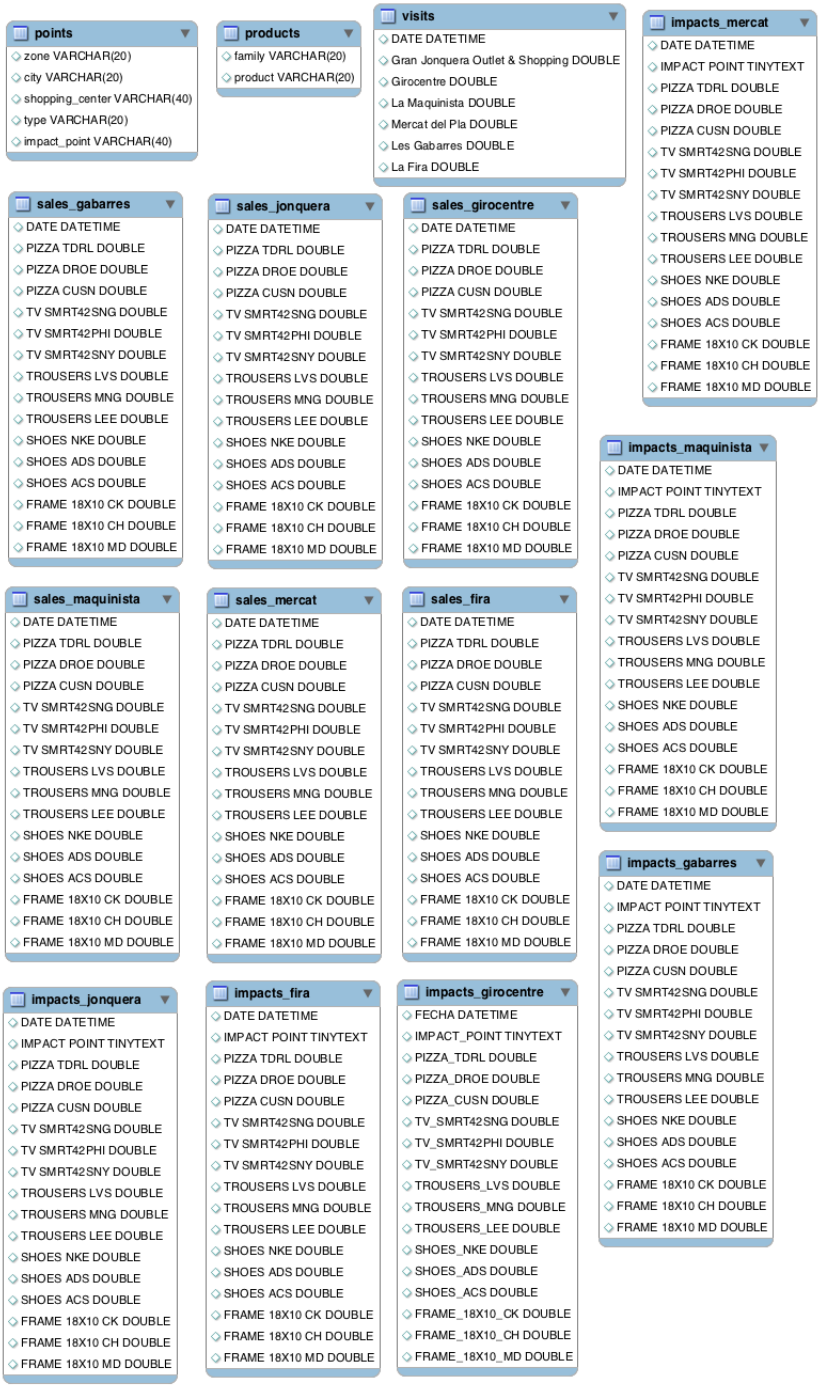



**Fig 4.** Model de dades original

Les taules anteriors ja disposen de totes les dades inicials necessàries per a la implementació del Data Warehouse, tot i així la estructuració de les taules no és la definitiva. Els camps de les taules, en canvi, gràcies a l'eina *Pentaho Data Integration* ja s'han creat amb el format correcte: *Strings, Date, Double*, etc.

### 3.3.2. Indicadors clau

Per poder aplicar les transformacions pertinents a les dades originals que ja tenim importades al *Data Warehouse*, es definiran els indicadors clau que ens guiaran en el procés de disseny i implementació del *Data Warehouse* final. Aquests indicadors claus s'extreuen dels requeriments i objectius del projecte.

Segons la definició del projecte, els indicadors claus són els següents:

- Impactes produïts al llarg del temps per un determinat article i punt publicitari
- Vendes diàries realitzades de cada article
- Nombre de visitants per dia als centres comercials

## 3.4. Model conceptual *Staging Area*

Tal i com s'ha descrit anteriorment per implementar el Data Warehouse definitiu, haurem d'aplicar un seguit de transformacions prèvies a l'aplicació processos ETL. Aquesta fase prèvia del model conceptual és l'anomenada *Staging Area.* Les transformacions afecten principalment a la distribució de les taules i relaciones entre elles mitjançant claus primàries (PK) i foranes (FK).

Implementar aquest model conceptual previ ens permetrà crear un sistema més escalable i per tant adaptable per si en un futur es vol afegir més centres comercials o productes sense necessitat de modificar el model conceptual de la base de dades, sempre que presentin el mateix format que les dades originals.
Seguidament definirem les transformacions que s'aplicaran al model conceptual inicial per establir els objectius esmentats amb l'objectiu d'aconseguir el model conceptual escalable.

1. Afegirem un nou camp ID que farà la funció de clau primària de la taula de punts d'impacte, anomenada ***points***. Aquest nou camp serà una

seqüència que anirà incrementant a mesura que s'introdueixen noves entrades a la taula.

També eliminarem el camp *impact point* ja que ja està present a la taula de *impacts*.

2. Afegirem un nou camp ID que farà la funció de clau primària de la taula de productes, anomenada ***products***. Aquest nou camp serà una seqüència que anirà incrementant a mesura que s'introdueixen noves entrades a la taula.

3. Unificació de les taules de vendes de cada un dels centres en una sola taula anomenada ***sales***. En aquesta taula la clau primària serà un nou camp *ID* i que serà una seqüència que anirà incrementant a mesura que s'introdueixen noves entrades a la taula. Aquesta taula també tindrà dos claus foranes. La primera serà sobre un nou camp ID_POINT referenciant el camp ID de la taula *points*. La segona clau forana serà sobre un camp nou ID_PRODUCT referenciant el camp ID de la taula *products*.

4. Unificació de les taules d'impactes de cada un dels punts d'impacte en una sola taula anomenada ***impacts***. En aquesta taula la clau primària serà un nou camp *ID* i que serà una seqüència que anirà incrementant a mesura que s'introdueixen noves entrades a la taula. Aquesta taula també tindrà dos claus foranes. La primera serà sobre un nou camp ID_POINT referenciant el camp ID de la taula *points*. La segona clau forana serà sobre un camp nou ID_PRODUCT referenciant el camp ID de la taula *products*.
5. Afegirem un camp ID a la taula ***visits***, aquest camp serà una seqüència que anirà incrementant a mesura que s'introdueixen noves entrades a la taula. Aquesta taula tindrà una clau forana, id_point, referenciant el camp ID de la taula *points*.

Després d'aplicar aquestes transformacions, el model conceptual queda definit segons la figura 5.


**Fig 5.** Model conceptual *Staging Area*

Aquesta transformació del model inicial i la càrrega de dades s'han realitzat directament amb dos Scripts SQL anomenats ***staging_area.sql*** i **loadStagingArea.sql** executats des de l'eina *MySQL Workbench*. Es poden veure aquests dos scripts a l'**Annex 1**.

## 3.5. Model conceptual del Data Warehouse

L'últim pas abans de explotar les dades amb les eines BI de *Pentaho*, és definir el model conceptual del *Data Warehouse* definitiu. Aquest model partirà de la definició de la *Staging Area* definida en l'apartat anterior. Un cop implementat

procedirem a carregar les dades mitjançant processos ETL, i seguidament es procedirà amb l'explotació i extracció d'informació sol·licitada a l'enunciat.

El model de dades que s'utilitzarà serà el **model multidimensional**. Aquest model és una tècnica de disseny lògic que busca presentar les dades en un estàndard, que permeti una recuperació adequada d'aquests. La idea fonamental és que l'usuari visualitzi fàcilment la relació que existeix entre els diferents components del model.

El model multidimensional és un model adequat que proveeix un camí viable per agregar ***fets*** al llarg de múltiples atributs, anomenats ***dimensions.*** Les dades son emmagatzemades com a ***fets*** i ***dimensions*** en un model de dades relacional.

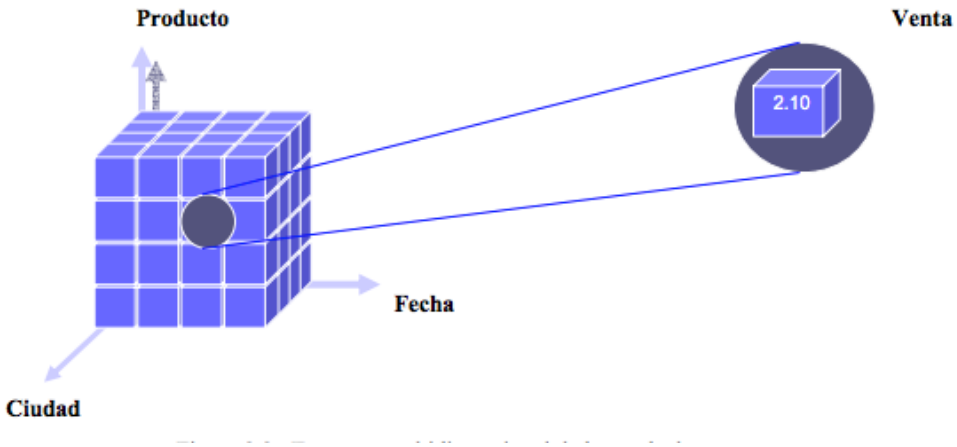



**Fig 6.** Esquema multidimensional de base de dades

### 3.5.1 Definicions del model multidimensional

A continuació es descriuran els diferents termes que composen el model multidimensional i que s'utilitzaran durant la definició del model del *Data Warehouse*.

- **Fets:** col·leccions de dades relacionades compostes per un indicador i un context. Les dimensions determinen els contexts dels fets, i cada fet particular està associat a un membre d'una dimensió.

- **Indicadors:** atributs numèrics associats als fets. Per exemple, número de visites, volum de vendes, etc.

- **Dimensions:** entitats respecte les quals el negoci vol mantenir organitzada la seva informació (productes, temps, clients, etc).

- **Membres:** noms o identificadors que marquen la seva posició dins d'una dimensió. Per exemple, NIF, mom, país, poden ser membres de la dimensió de clients.

- **Jerarquies:** Els membres de les dimensions es solen organitzar de forma jeràrquica.

### 3.5.2 Dimensions i fets

Una vegada definits els conceptes principals del model multidimensional, es definiran a continuació les *dimensions, jerarquies associades, fets, indicadors i mesures* que es necessiten per la implementació del model. El model escollit serà de tipus **estrella**, o les taules estan unides a la taula central a través de les seves respectius claus.

**DIMENSIONS**

| **Nom** | **POINTS** |
|---|---|
| **Descripció** | Conté els 6 punts d'impacte, és a dir, els centres comercials que s'estan analitzant. |
| **Atributs** | • ID (PK)<br>• Zona<br>• Ciutat<br>• Nom<br>• Tipus |
| **Jerarquia** | Nivell 0: Zona |

| | |
|---|---|
| | Nivell 1: Tipus<br>Nivell 2: Ciutat<br>Nivell 3: Centre comercial |

| **Nom** | **PRODUCTS** |
|---|---|
| **Descripció** | Conté els 16 productes de l'estudi |
| **Atributs** | • ID (PK)<br>• Família<br>• Producte |
| **Jerarquia** | Nivell 0: Família<br>Nivell 1: Producte |

| **Nom** | **TEMPS** |
|---|---|
| **Descripció** | Conté l'atribut temps en els seus diferents nivells jeràrquics |
| **Atributs** | • ID (PK)<br>• Data<br>• Any<br>• Estació de l'any<br>• Trimestre<br>• Mes<br>• Nom del mes<br>• Dia del mes<br>• Nom del dia |
| **Jerarquia** | Nivell 0: Any<br>Nivell 1: Estació de l'any<br>Nivell 2: Trimestre<br>Nivell 3: Mes<br>Nivell 4: Dia |

La dimensió ***temps*** modela el moment en el qual es produeixen tots els *fets* que es detallaran a continuació. En el model inicial aquest atribut era en format *datetime* a les taules de: *sales, impacts i visits*. En el model multidimensional es crea aquesta *dimensió* per emmagatzemar la informació temporal que afecta

als *fets* esmentats. Això és necessari ja que el anàlisis de la informació requereix l'estudi temporal amb diferents nivells.

**FETS**

| **Nom** | **SALES_IMPACTS** |
|---|---|
| **Descripció** | Conté el registre de les vendes i impactes produïts en cada un dels punts i per cada producte. A través dels indicadors corresponents reportarà informació sobre els productes venuts, els impactes produïts i la distribució temporal i per centres comercials d'aquestes vendes. |
| **Atributs** | • ID<br>• ID producte<br>• ID centre<br>• ID del temps<br>• Vendes<br>• Impactes Big Screen<br>• Impactes Information Point Screen |
| **Dimensions** | • Producte<br>• Centre<br>• Temps |
| **Indicadors** | • Nombre de vendes de cada producte a cadascun dels punts.<br>• Nombre de vendes en cada moment temporal.<br>• Nombre de vendes de cada família de productes.<br>• Nombre de vendes per zona del centre.<br>• Nombre d'impactes de cada producte a cadascun dels punts.<br>• Nombre d'impactes en cada moment temporal.<br>• Nombre d'impactes de cada família de productes.<br>• Nombre d'impactes per zona geogràfica.<br>• Nombre d'impactes per tipus de punt d'impacte. |

| **Nom** | **VISITS** |
|---|---|
| **Descripció** | Conté el registre de les visites en cada un dels punts. A través dels indicadors corresponents reportarà informació sobre el nombre de visites rebudes i la distribució temporal i per centres comercials d'aquestes visites. |
| **Atributs** | • ID centre<br>• ID del temps<br>• Visites |
| **Dimensions** | • Centre<br>• Temps |
| **Indicadors** | • Nombre de visites a cadascun dels punts.<br>• Nombre de visites en cada moment temporal.<br>• Nombre de visites per zona geogràfica |

Com que a les taules de *sales* i *impacts* tenim informació del producte, podem ajuntar-les en una sola taula de *fets*. D'aquesta manera tindrem en un sol cub la informació de ventes i impactes i ens permetrà tenir millors opcions d'anàlisis en un futur.

### 3.5.3 Diagrama del model multidimensional

Després de definir els fets i dimensions junt amb les jerarquies i indicadors, aquest és el diagrama del model multidimensional resultant:

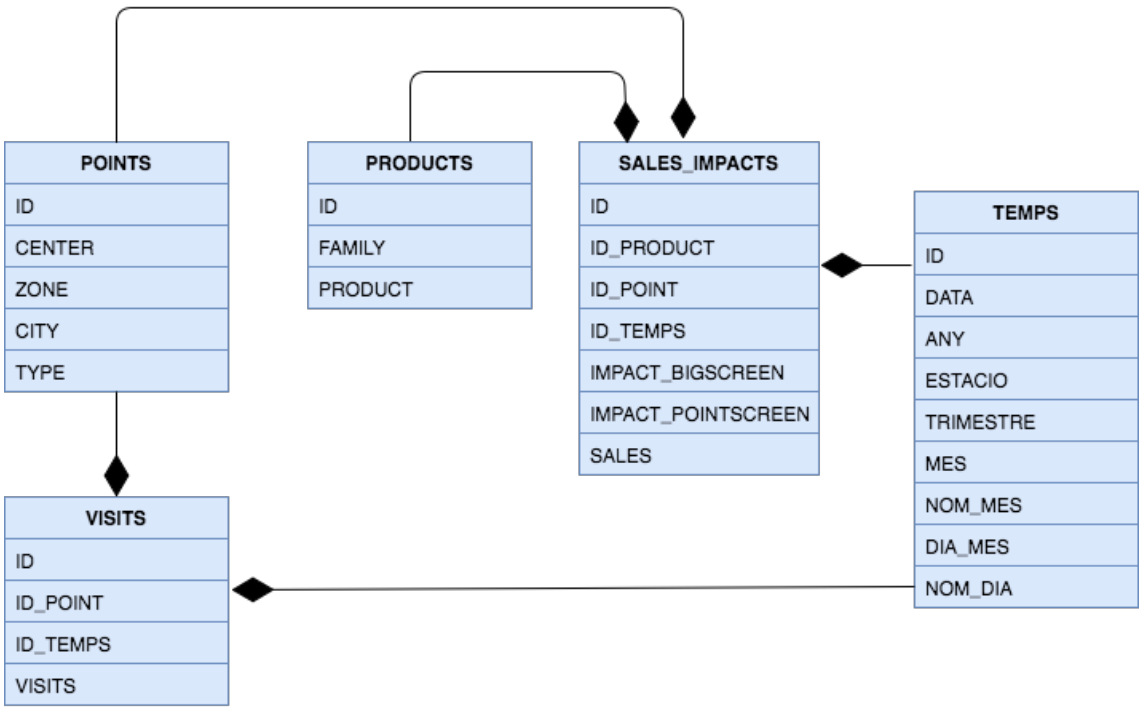



**Fig 7.** Model multidimensional conceptual del *Data Warehouse*

## 3.6 Implementació del *Data Warehouse*

Per la implementació final del *Data Warehouse* crearem un esquema nou a la base de dades amb les *dimensions* i taules de *fets*. Aquesta base de dades es crearà i omplirà mitjançant processos ETL a través de processos SQL. Aquests processos SQL ens permetran realitzar dos tasques, la creació de les *dimensions* i taules de *fets* en primer lloc, i en segon omplir-les de les dades, incloent els diferents càlculs dels atributs de les taules de *fets*.

### 3.6.1 Taules de la base de dades "PubliDW"

La figura següent mostra el diagrama de les taules que corresponen a la implementació del *Data Warehouse* modelat a l'apartat 3.5.3. Aquest diagrama conté les relacions entre les *dimensions* i taules de *fets* del model conceptual. També mostra els atributs que identifiquen els registres, les relacions entre taules i el format dels diferents atributs.

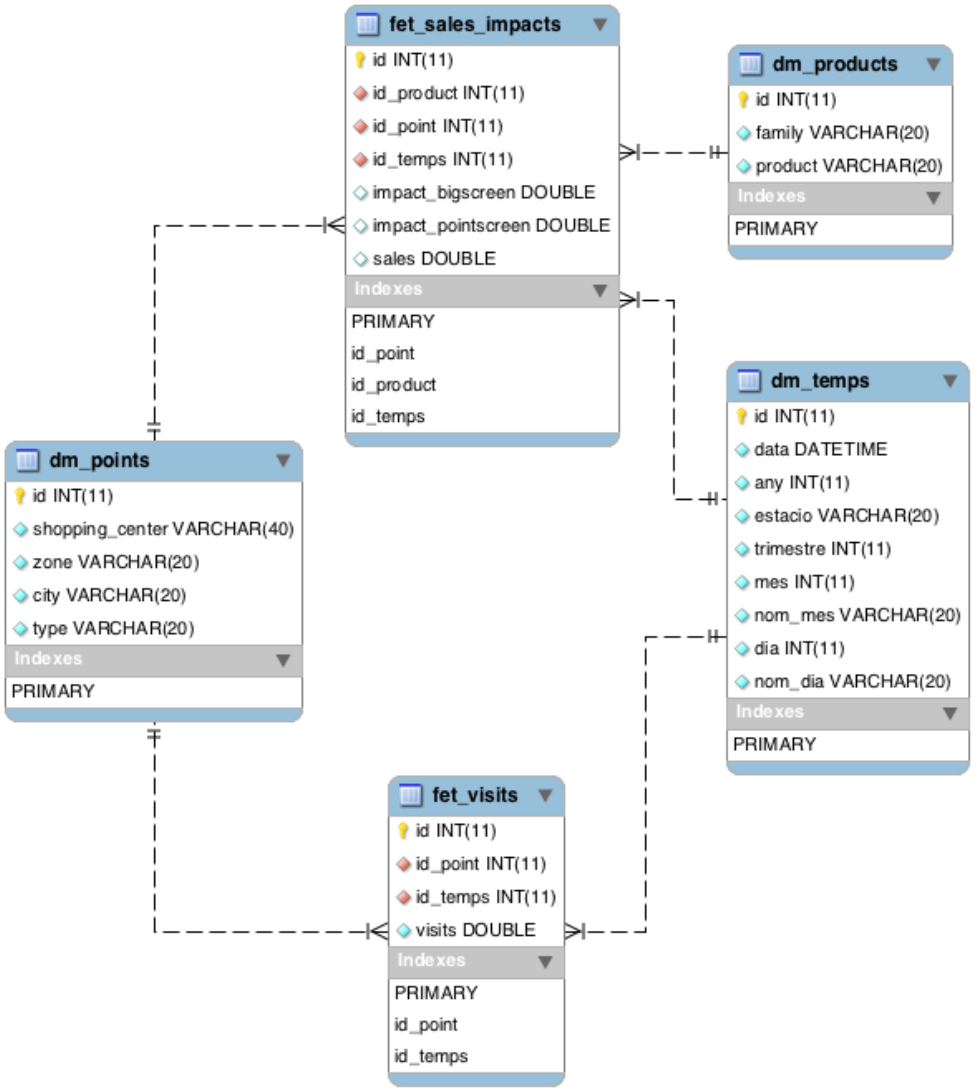



**Fig 8.** Model del Data Warehouse PubliDW

A l'hora de definir el *Data Warehouse PubliDW* hi ha un seguit de punts a destacar envers la definició anterior del *Staging Area*:

1. Dins de la dimensió *temps* s'han afegit una sèrie d'atributs per ampliar el detall d'informació proporcionada pels camps de format *datetime* a les taules de *visits, sales i impacts*:

    a. ***data*:** Camp de tipus *datetime* que guarda la data completa.

b. ***any:*** Camp de tipus *Integer* que guarda els quatre dígits de l'any: 2016.
c. ***estacio:*** Camp de tipus *varchar* que guarda l'estació de l'any: primavera, estiu, tardor, hivern.
d. ***trimestre:*** Camp de tipus *Integer* que guarda el trimestre*: 1, 2, 3.*
e. ***mes:*** Camp de tipus *Integer* que guarda el valor numèric del mes: 1-12.
f. ***nom_mes:*** Camp de tipus *varchar* que guarda el nom del mes.
g. ***dia:*** Camp de tipus *Integer* que guarda el número del dia..
h. ***nom_dia:*** Camp de tipus *varchar* que guarda el nom del dia.

2. S'han unit les taules *impacts* i *sales* del model de la *Staging Area* ja que les dues taules tenen dependència de les taules de *points* i *products*. Cada registre de la nova taula de *fet_sales_impacts* contendrà el valor de les ventes i els dos tipus d'impactes: *Big Screen i Information Point Screen*. Aquesta unió ens permetrà tenir els tres valors en un mateix cub OLAP, i per tant, millors opcions d'anàlisis.

## 3.7 Càrrega de dades al *Data Warehouse*

A continuació s'exposen els scripts SQL i *stored procedures* que s'encarreguen d'anar omplint la base de dades ***PubliDW*** a partir de la ***StagingArea***, i de fer els càlculs corresponents per tal d'omplir degudament la taula de dimensions *temps*, que és la única que requereix de certs càlculs previs per tal de confeccionar-la amb les dades que es necessiten.

**Script loadPubliDW.sql**

Aquest script carrega les dades de la base de dades Staging Area a les taules de dimensions i fets de la PubliDW.

```
SET FOREIGN_KEY_CHECKS = 0;
truncate table publidw.fet_sales_impacts;
truncate table publidw.fet_visits;
truncate table publidw.dm_points;
truncate table publidw.dm_products;
truncate table publidw.dm_temps;
SET FOREIGN_KEY_CHECKS = 1;

-- ----------------------------------------------------
-- Insert `publidw`.`dm_points`
-- ----------------------------------------------------
INSERT INTO publidw.dm_points (id,zone,city,shopping_center, type)
SELECT id,zone,city,shopping_center,type from staging_area.points;

-- ----------------------------------------------------
-- Insert `publidw`.`dm_products`
-- ----------------------------------------------------
INSERT INTO publidw.dm_products (id,family,product)
SELECT id,family,product from staging_area.products;
```

```
-- ----------------------------------------------------
-- Insert `publidw`.`dm_temps`
-- ----------------------------------------------------
INSERT INTO publidw.dm_temps (data,any,estacio,trimestre,mes,nom_mes,dia,nom_dia)
SELECT visits_date, EXTRACT(YEAR FROM visits_date),
(CASE WHEN (EXTRACT(MONTH FROM visits_date) >= 3 and EXTRACT(DAY FROM visits_date) >= 1)
and  (EXTRACT(MONTH FROM visits_date) <= 5 and EXTRACT(DAY FROM visits_date) <= 31)
THEN 'Primavera'
WHEN (EXTRACT(MONTH FROM visits_date) >= 6 and EXTRACT(DAY FROM visits_date) >= 1)
and  (EXTRACT(MONTH FROM visits_date) <= 8 and EXTRACT(DAY FROM visits_date) <= 31)
THEN 'Estiu'
WHEN (EXTRACT(MONTH FROM visits_date) >= 9 and EXTRACT(DAY FROM visits_date) >= 1)
and  (EXTRACT(MONTH FROM visits_date) <= 11 and EXTRACT(DAY FROM visits_date) <= 31)
THEN 'Tardor'
WHEN (EXTRACT(MONTH FROM visits_date) >= 12 and EXTRACT(DAY FROM visits_date) >= 1)
and  (EXTRACT(MONTH FROM visits_date) <= 12 and EXTRACT(DAY FROM visits_date) <= 31)
THEN 'Hivern'
WHEN (EXTRACT(MONTH FROM visits_date) >= 1 and EXTRACT(DAY FROM visits_date) >= 1)
and  (EXTRACT(MONTH FROM visits_date) <= 2 and EXTRACT(DAY FROM visits_date) <= 31)
THEN 'Hivern' ELSE 'ND'
END) as estacio,
(CASE
  WHEN MONTH(visits_date) IN (1,2,3,4) THEN 1
  WHEN MONTH(visits_date) IN (5,6, 7,8) THEN 2
  WHEN MONTH(visits_date) IN (9, 10,11,12) THEN 3
  END) as trimestre,
EXTRACT(MONTH FROM visits_date) as mes,
MONTHNAME(visits_date) as nom_mes,
EXTRACT(DAY FROM visits_date) as dia,
DAYNAME(visits_date) as nom_dia
from staging_area.visits group by (visits_date);
```

```
-- -----------------------------------------------------
-- Insert `publidw`.`fet_visits`
-- -----------------------------------------------------
INSERT INTO publidw.fet_visits (id_point, id_temps, visits)
SELECT p.id, t.id, v.visits from staging_area.visits v left join publidw.dm_temps t
on v.visits_date = t.data left join publidw.dm_points p on p.id = v.id_point;

-- -----------------------------------------------------
-- Insert `publidw`.`fet_sales_impacts`
-- -----------------------------------------------------
INSERT INTO publidw.fet_sales_impacts (id_product,id_point,id_temps, sales)
SELECT pr.id, p.id, t.id,sales from staging_area.sales s
left join publidw.dm_temps t on s.sale_date = t.data left join publidw.dm_points p on p.id = s.id_point
left join publidw.dm_products pr on pr.id = s.id_product;

SET SQL_SAFE_UPDATES = 0;
UPDATE publidw.fet_sales_impacts fi left join publidw.dm_temps t on t.id = fi.id_temps
left join staging_area.impacts si on fi.id_point = si.id_point
and fi.id_product = si.id_product and t.data = si.impact_date
SET fi.impact_bigscreen = si.impacts
WHERE si.impact_point = 'Big Screen';

UPDATE publidw.fet_sales_impacts fi left join publidw.dm_temps t on t.id = fi.id_temps
left join staging_area.impacts si on fi.id_point = si.id_point
and fi.id_product = si.id_product and t.data = si.impact_date
SET fi.impact_pointscreen = si.impacts
WHERE si.impact_point = 'Information Point Screen';
```

### 3.7.1 Estat de les taules definitives *Data Warehouse*

Un cop executats els scripts de creació i càrrega anteriors ja tenim el *Data Warehouse* preparat per començar a realitzar l'anàlisi objectiu del treball.

Les figures següents mostren l'estat de les taules del *PubliDW* des de l'eina *MySQL Workbench.*

| id | shopping_center | zone | city | type |
|---|---|---|---|---|
| 1 | Gran Jonquera Outlet & Shopping | NORTH | Figueres | Periphery |
| 2 | Girocentre | NORTH | Girona | Urban |
| 3 | La Maquinista | CENTER | Barcelona | Urban |
| 4 | Mercat del Pla | CENTER | Lleida | Urban |
| 5 | Les Gabarres | SOUTH | Tarragona | Periphery |
| 6 | La Fira | SOUTH | Reus | Periphery |

**Fig 9.** Contingut de la taula *dm_points*

| id | family | product |
|---|---|---|
| 1 | Food | PIZZA TDRL |
| 2 | Food | PIZZA DROE |
| 3 | Food | PIZZA CUSN |
| 4 | Electronics | TV SMRT42SNG |
| 5 | Electronics | TV SMRT42PHI |
| 6 | Electronics | TV SMRT42SNY |
| 7 | Clothing | TROUSERS LVS |
| 8 | Clothing | TROUSERS MNG |
| 9 | Clothing | TROUSERS LEE |
| 10 | Sports | SHOES NKE |
| 11 | Sports | SHOES ADS |
| 12 | Sports | SHOES ACS |
| 13 | Home | FRAME 18X10 CK |
| 14 | Home | FRAME 18X10 CH |
| 15 | Food | FRAME 18X10 MD |

**Fig 10.** Contingut de la taula *dm_products*

| id | data | any | estacio | trimestre | mes | nom_mes | dia | nom_dia |
|---|---|---|---|---|---|---|---|---|
| 48 | 2016-02-17 00:00:00 | 2016 | Hivern | 1 | 2 | February | 17 | Wednesday |
| 49 | 2016-02-18 00:00:00 | 2016 | Hivern | 1 | 2 | February | 18 | Thursday |
| 50 | 2016-02-19 00:00:00 | 2016 | Hivern | 1 | 2 | February | 19 | Friday |
| 51 | 2016-02-20 00:00:00 | 2016 | Hivern | 1 | 2 | February | 20 | Saturday |
| 52 | 2016-02-21 00:00:00 | 2016 | Hivern | 1 | 2 | February | 21 | Sunday |
| 53 | 2016-02-22 00:00:00 | 2016 | Hivern | 1 | 2 | February | 22 | Monday |
| 54 | 2016-02-23 00:00:00 | 2016 | Hivern | 1 | 2 | February | 23 | Tuesday |
| 55 | 2016-02-24 00:00:00 | 2016 | Hivern | 1 | 2 | February | 24 | Wednesday |
| 56 | 2016-02-25 00:00:00 | 2016 | Hivern | 1 | 2 | February | 25 | Thursday |
| 57 | 2016-02-26 00:00:00 | 2016 | Hivern | 1 | 2 | February | 26 | Friday |
| 58 | 2016-02-27 00:00:00 | 2016 | Hivern | 1 | 2 | February | 27 | Saturday |
| 59 | 2016-02-28 00:00:00 | 2016 | Hivern | 1 | 2 | February | 28 | Sunday |
| 60 | 2016-02-29 00:00:00 | 2016 | Hivern | 1 | 2 | February | 29 | Monday |
| 61 | 2016-03-01 00:00:00 | 2016 | Primavera | 1 | 3 | March | 1 | Tuesday |
| 62 | 2016-03-02 00:00:00 | 2016 | Primavera | 1 | 3 | March | 2 | Wednesday |
| 63 | 2016-03-03 00:00:00 | 2016 | Primavera | 1 | 3 | March | 3 | Thursday |
| 64 | 2016-03-04 00:00:00 | 2016 | Primavera | 1 | 3 | March | 4 | Friday |
| 65 | 2016-03-05 00:00:00 | 2016 | Primavera | 1 | 3 | March | 5 | Saturday |
| 66 | 2016-03-06 00:00:00 | 2016 | Primavera | 1 | 3 | March | 6 | Sunday |
| 67 | 2016-03-07 00:00:00 | 2016 | Primavera | 1 | 3 | March | 7 | Monday |
| 68 | 2016-03-08 00:00:00 | 2016 | Primavera | 1 | 3 | March | 8 | Tuesday |
| 69 | 2016-03-09 00:00:00 | 2016 | Primavera | 1 | 3 | March | 9 | Wednesday |
| 70 | 2016-03-10 00:00:00 | 2016 | Primavera | 1 | 3 | March | 10 | Thursday |

**Fig 11.** Contingut de la taula *dm_temps*

| id | id_point | id_temps | visits |
|---|---|---|---|
| 4 | 1 | 1 | 13757 |
| 5 | 2 | 1 | 12839 |
| 6 | 3 | 1 | 20546 |
| 7 | 4 | 1 | 10360 |
| 8 | 5 | 1 | 22286 |
| 9 | 6 | 1 | 11420 |
| 10 | 1 | 2 | 0 |
| 11 | 2 | 2 | 0 |
| 12 | 3 | 2 | 0 |
| 13 | 4 | 2 | 0 |
| 14 | 5 | 2 | 0 |
| 15 | 6 | 2 | 0 |
| 16 | 1 | 3 | 9218 |
| 17 | 2 | 3 | 8957 |
| 18 | 3 | 3 | 14758 |
| 19 | 4 | 3 | 5549 |
| 20 | 5 | 3 | 10943 |
| 21 | 6 | 3 | 7782 |
| 22 | 1 | 4 | 6869 |
| 23 | 2 | 4 | 9742 |
| 24 | 3 | 4 | 13976 |
| 25 | 4 | 4 | 7339 |
| 26 | 5 | 4 | 11035 |
| 27 | 6 | 4 | 7410 |
| 28 | 1 | 5 | 9167 |
| 29 | 2 | 5 | 6784 |

**Fig 12.** Contingut de la taula *fet_visits*

| id | id_product | id_point | id_temps | impact_bigscre... | impact_pointscre... | sales |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 2 | 1 | 83 |
| 2 | 2 | 1 | 1 | 7 | 4 | 291 |
| 3 | 3 | 1 | 1 | 9 | 9 | 443 |
| 4 | 4 | 1 | 1 | 1 | 2 | 3385 |
| 5 | 5 | 1 | 1 | 6 | 5 | 3667 |
| 6 | 6 | 1 | 1 | 12 | 11 | 4476 |
| 7 | 7 | 1 | 1 | 1 | 3 | 159 |
| 8 | 8 | 1 | 1 | 8 | 7 | 248 |
| 9 | 9 | 1 | 1 | 10 | 9 | 259 |
| 10 | 10 | 1 | 1 | 2 | 3 | 552 |
| 11 | 11 | 1 | 1 | 4 | 7 | 1038 |
| 12 | 12 | 1 | 1 | 11 | 11 | 2292 |
| 13 | 13 | 1 | 1 | 1 | 2 | 236 |
| 14 | 14 | 1 | 1 | 5 | 4 | 229 |
| 15 | 15 | 1 | 1 | 10 | 11 | 295 |
| 16 | 1 | 2 | 1 | 1 | 2 | 83 |
| 17 | 2 | 2 | 1 | 5 | 7 | 110 |
| 18 | 3 | 2 | 1 | 9 | 11 | 101 |
| 19 | 4 | 2 | 1 | 2 | 2 | 1500 |
| 20 | 5 | 2 | 1 | 6 | 4 | 4020 |
| 21 | 6 | 2 | 1 | 9 | 10 | 6540 |
| 22 | 7 | 2 | 1 | 3 | 2 | 282 |
| 23 | 8 | 2 | 1 | 8 | 5 | 462 |
| 24 | 9 | 2 | 1 | 11 | 9 | 996 |
| 25 | 10 | 2 | 1 | 3 | 3 | 362 |
| 26 | 11 | 2 | 1 | 5 | 6 | 304 |
| 27 | 12 | 2 | 1 | 9 | 12 | 370 |
| 28 | 13 | 2 | 1 | 1 | 3 | 200 |
| 29 | 14 | 2 | 1 | 4 | 8 | 214 |

**Fig 13.** Contingut de la taula fet_sales_impacts

# 4. Explotació de Dades

## 4.1 Creació dels cubs OLAP amb *Pentaho Schema Workbench*

L'explotació de dades d'un *Data Warehouse* comença amb la creació de cubs OLAP (*Online Analytical Processing)*. La definició del model multidimensional de l'apartat anterior ens permetrà definir i dissenyar aquests cubs. L'eina utilitzada per aquesta tasca és *Pentaho Schema Workbench,* que permet crear una connexió amb la nostra base de dades i a partir d'aquí poder confeccionar els diferents cubs i els objectes que el formen: dimensions, jerarquies, nivells i mesures. Finalment aquesta eina ens permetrà publicar a *Pentaho BI Server* els cubs creats per al seu anàlisis.

Per tal que els cubs es puguin publicar correctament s'han de complir les següents restriccions:

- Cada cub ha de tenir definit al menys una mesura i ha d'estar relacionat amb una taula de fets i amb una dimensió com a mínim.
- Cada dimensió ha de tenir una jerarquia com a mínim.
- Cada jerarquia ha d'estar relacionada amb una taula de dimensions.
- Cada nivell d'una jerarquia ha d'especificar la columna de la taula relacionada amb la seva jerarquia, el tipus de dades i el tipus d'atribut del nivell dins de la jerarquia.

La connexió a la base de dades es realitza des del menú *Options/Connection* i funciona de la mateixa manera que en *Pentaho Data Integration*. Un cop establerta connexió entre la base de dades i *Schema Workbench* ja es poden crear els cubs OLAP. La creació d'aquests cubs és molt senzilla i l'aplicació en tot moment indica a l'usuari els problemes de definició que van apareixen al llarg del procés.

Les següents figures mostren els dos cubs creats.

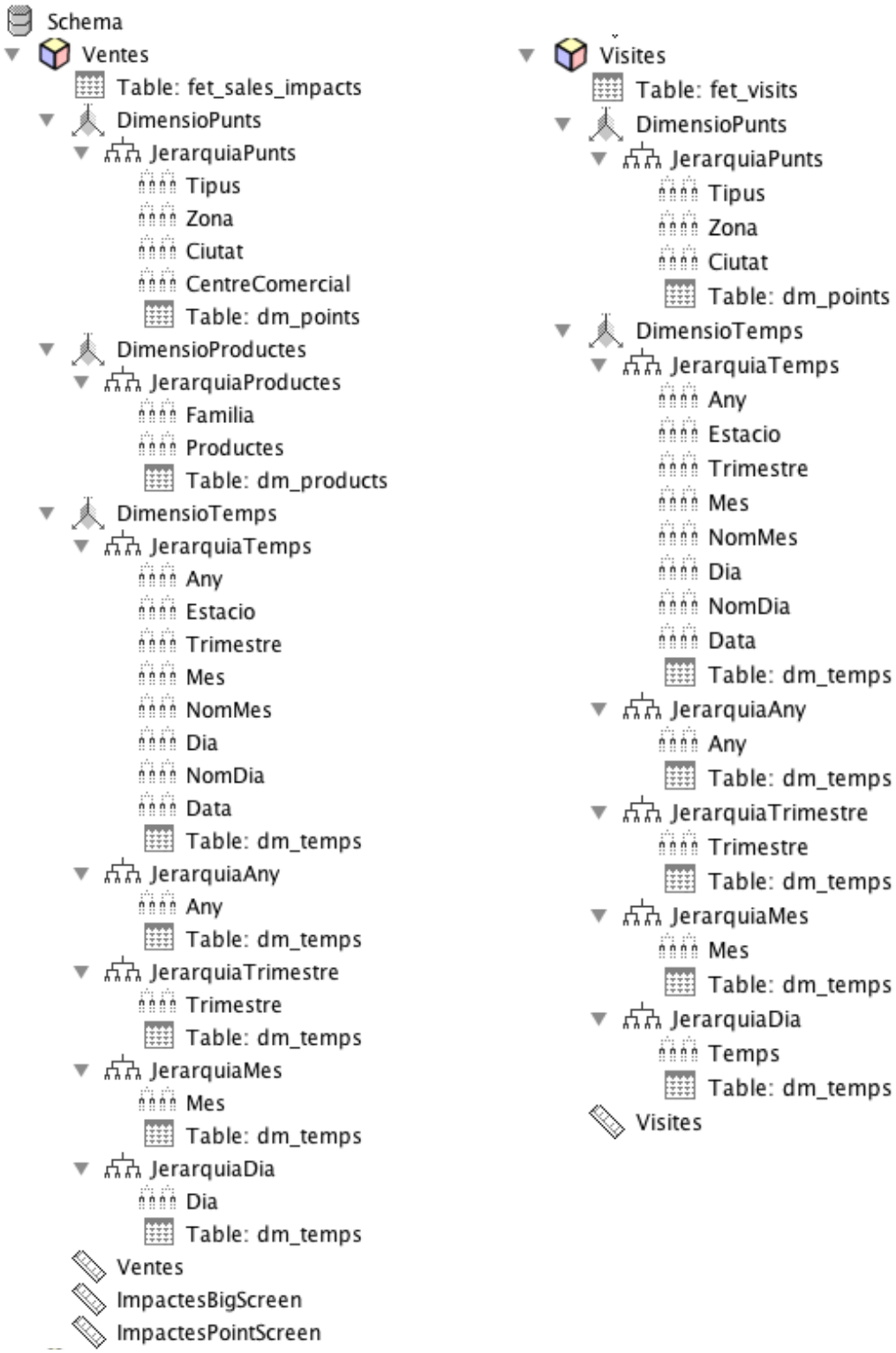



**Fig 14.** Estructura dels cubs OLAP: Ventes (esquerra) i Visites (dreta)

Com podem observar a la figura anterior s'han creat dos cubs per a cada una de les taules de fets: *fet_sales_impacts (cub ventes)* i fet*_visits (cub visites).* Pel que fa al cub de Ventes s'han creat tres dimensions *DimensióPunts, DimensióProductes i DimensióTemps.* Pel cub de Visites s'han creat les mateixes dimensions excepte la *DimensióProductes.* Per cada una de les diferents dimensions s'han definit les seves jerarquies tal com s'havien especificat a l'apartat 3.5.2.

La següent taula ens mostra la distribució de les dimensions, jerarquies i nivells que es poden utilitzar en els filtres dels cubs de Ventes i Visites.

| Cub ***Ventes*** | | | Cub ***Visites*** | | |
|---|---|---|---|---|---|
| **Indicador** | *Ventes*<br>*Impactes Big Screen*<br>*Impactes Point Screen* | | **Indicador** | *Nombre Visites* | |
| **Dimensió** | **Jerarquia** | **Nivell** | **Dimensió** | **Jerarquia** | **Nivell** |
| DimensioPunts | JerarquiaPunts | Tipus<br>Zona<br>Ciutat<br>Centre | DimensiPunts | JerarquiaPunts | Tipus<br>Zona<br>Ciutat<br>Centre |
| DimensioProductes | JerarquiaProductes | Familia<br>Productes | | | |
| DimensioTemps | JerarquiaTemps | Any<br>Estacio<br>Trimestre<br>Mes<br>NomMes<br>Dia<br>NomDia<br>Data | DimensioTemps | JerarquiaTemps | Any<br>Estacio<br>Trimestre<br>Mes<br>NomMes<br>Dia<br>NomDia<br>Data |
| | JerarquiaAny | Any | | JerarquiaAny | Any |
| | JerarquiaTrimes | Trimestre | | JerarquiaTrimes | Trimestre |
| | JerarquiaMes | Mes | | JerarquiaMes | Mes |
| | JerarquiaDia | Dia | | JerarquiaDia | Dia |

Un cop creats els cubs amb l'eina *Pentaho Schema Workbench,* podrem obtenir el fitxer ***CubPubliDW.xml****,* que contindrà la informació referent a la configuració del cub però sense els paràmetres de la connexió a la base de dades. Aquest fitxer es pot veure a l'**Annex 2.**

Els nivells i dimensions assignats a cadascun dels dos cubs implementats no són imprescindibles per respondre les qüestions plantejades en els objectius d'aquest projecte. L'objectiu és crear una estructura OLAP que ens permeti tenir un ampli ventall de possibilitats d'anàlisis a realitzar.

## 4.2 Navegació cubs OLAP amb Jpivot

*Jpivot* és un conjunt de llibreries customitzables JSP que permet navegar cubs OLAP i mostrar la informació en forma de taules i gràfics, suportant la funcionalitat típica dels entorns OLAP com *drill-down*, rotar eixos, entre altres.

Dins de *Pentaho BI Server CE* ve instal·lada l'eina *Jpivot* que ens permetrà fer un primer anàlisi dels cubs per comprovar que les dades que contenen són correctes i que les dimensions, jerarquies i nivells són suficients per donar resposta als objectius d'aquest projecte.

A continuació es mostren els passos necessaris per analitzar els cubs publicats des de *Pentaho Scehma Workbench* amb *JPivot.*

1. Descarregar *Pentaho Server* i desplegar el *Tomcat*.
2. Introduir l'adreça *http://localhost:8080* al navegador web.

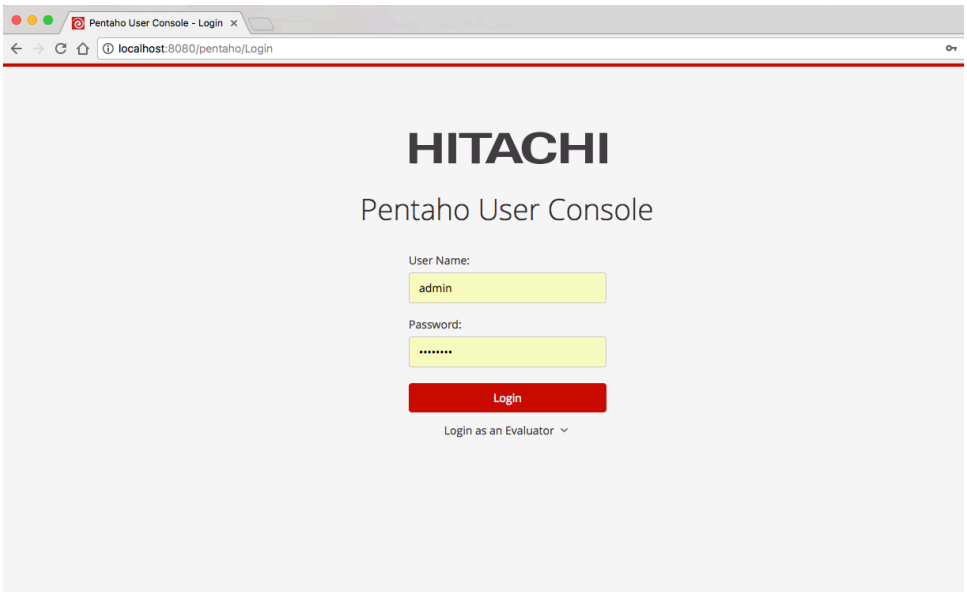



**Fig 15.** Login *Pentaho Server CE*

3. Afegir el *Datasource* i la connexió a la base de dades.


**Fig 16.** Gestió *Datasource* de *Pentaho Server*

4. Publicar el cub des de *Pentaho Schema Workbench* assegurant-nos de que el nom del *Datasource* coincideixi amb el de *Pentaho Server*.

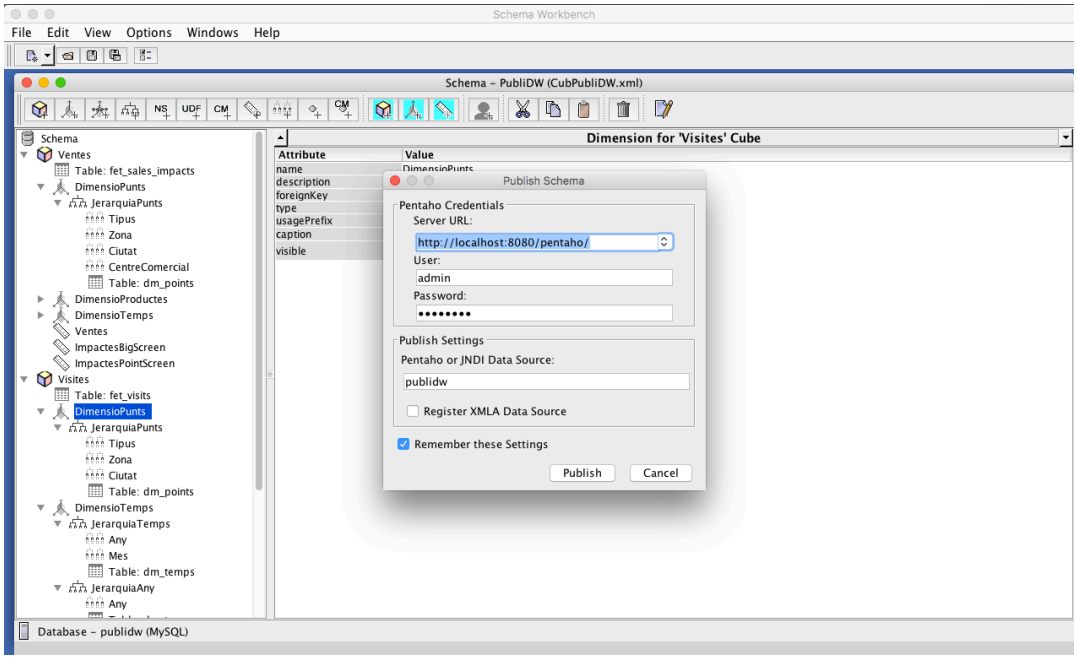



**Fig 17.** Publicació del esquema des de *Schema Workbench*

5. Crear una nova vista *JPivot* i seleccionar l'esquema que es correspongui amb el nom que s'ha publicat des de *Pentaho Schema Workbench* i el cub que es vulgui analitzar de l'esquema seleccionat.

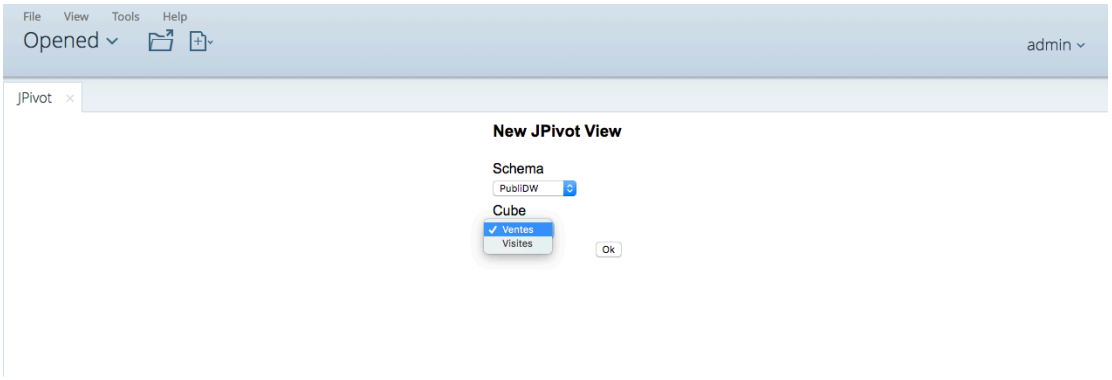
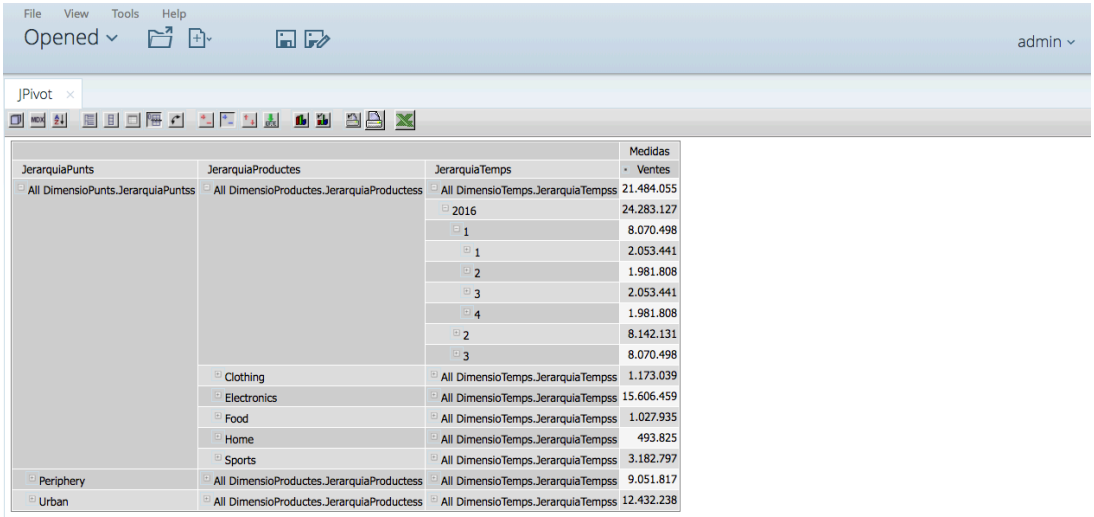



**Fig 18.** Nova vista *JPivot*

6. A continuació s'obrirà l'estructura de dades del cub OLAP amb tots els seus objectes en forma de desplegables. L'eina ens permet navegar pel cub interactuant amb les jerarquies i nivells de les diferents dimensions.




| | | | Medidas |
|---|---|---|---|
| JerarquiaPunts | JerarquiaProductes | JerarquiaTemps | Ventes |
| All DimensioPunts.JerarquiaPuntss | All DimensioProductes.JerarquiaProductess | All DimensioTemps.JerarquiaTempss | 21.484.055 |
| | | 2016 | 24.283.127 |
| | | 1 | 8.070.498 |
| | | 1 | 2.053.441 |
| | | 2 | 1.981.808 |
| | | 3 | 2.053.441 |
| | | 4 | 1.981.808 |
| | | 2 | 8.142.131 |
| | | 3 | 8.070.498 |
| | Clothing | All DimensioTemps.JerarquiaTempss | 1.173.039 |
| | Electronics | All DimensioTemps.JerarquiaTempss | 15.606.459 |
| | Food | All DimensioTemps.JerarquiaTempss | 1.027.935 |
| | Home | All DimensioTemps.JerarquiaTempss | 493.825 |
| | Sports | All DimensioTemps.JerarquiaTempss | 3.182.797 |
| Periphery | All DimensioProductes.JerarquiaProductess | All DimensioTemps.JerarquiaTempss | 9.051.817 |
| Urban | All DimensioProductes.JerarquiaProductess | All DimensioTemps.JerarquiaTempss | 12.432.238 |

**Fig 19.** Análisis del cub a través de *JPivot*

Com es pot comprovar *JPivot* ens permet explorar les dades i comprovar que l'estructura de cubs OLAP és correcta en quan a la seva sintaxis, i respecta al contingut si les jerarquies i nivells de les dimensions corresponen a les dissenyades des del *Schema Workbench.* No obstant, aquesta no és la millor

eina per tal de començar a explotar les dades del cub i fer-ne els anàlisis corresponents. Per això hi ha altres eines molt més complexes i eficients que serveixen per analitzar en profunditat les dades i generar-ne tots els gràfics i informes necessaris.

En aquest projecte s'ha optat per utilitzar el *plugin* ***Saiku Analytics***, una eina que ens permetrà generar els anàlisis de les ventes i impactes per productes dels centres comercials.

## 4.3 Navegació i explotació dels cubs de dades amb *Saiku Analytics*

*Saiku* és una excel·lent visor OLAP que proporciona al usuari una magnífica eina per analitzar de forma fàcil i intuïtiva.

A continuació es detallen els passos per instal·lar *Saiku Analytics:*

1. Des de *Pentaho Server* seleccionem l'opció *Marketplace* des del desplegable d'opcions del menú *Home*.

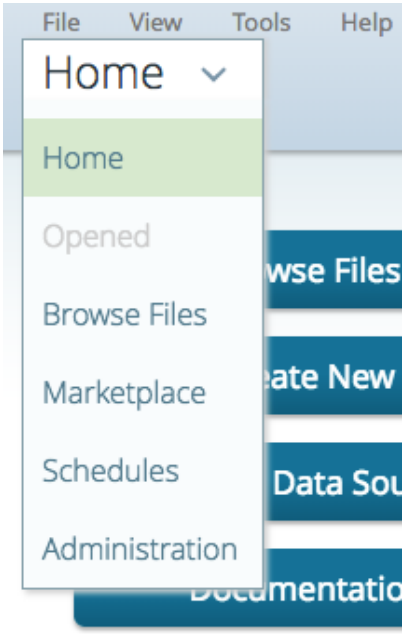



**Fig 20.** Menú *Pentaho Server*

2. Cercar *Saiku* i instal·lar el *plugin Saiku Analytics*.

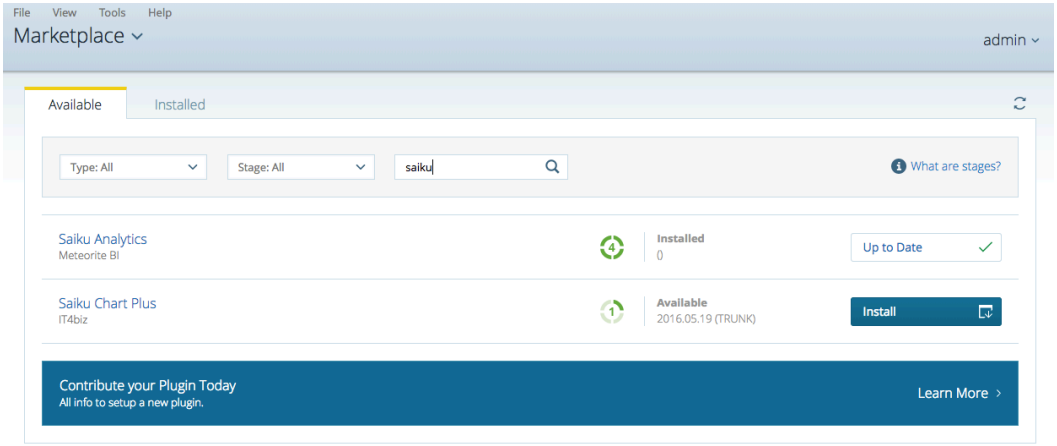



**Fig 21.** Marketplace *Pentaho Server*

3. Reiniciar *Pentaho Server* per poder utilitzar *Saiku*.
4. Si s'ha instal·lat correctament, podrem seleccionar *Saiku Analytics* des del menú.

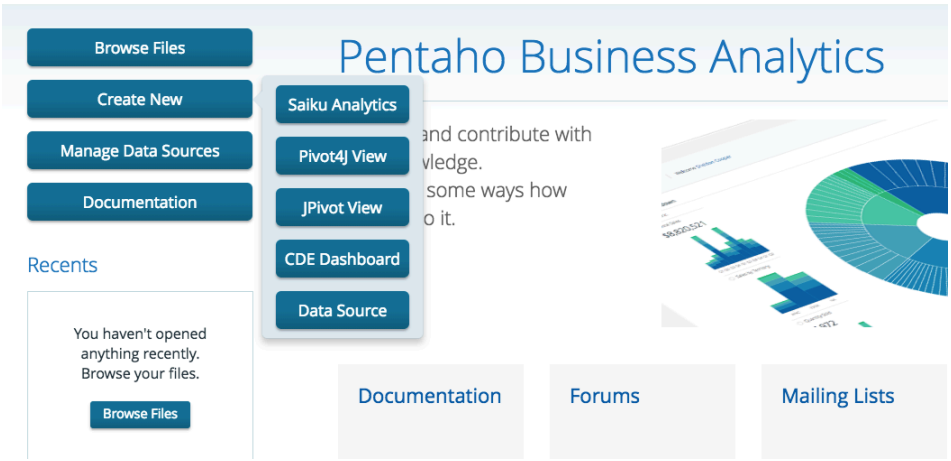



**Fig 22.** Menú *Create New* de *Pentaho Server*

Un cop instal·lat el *plugin* es detallaran els passos per realitzar un anàlisis amb *Saiku Analytics.*

1. Accedir al *plugin* des de ***Create New/Saiku Analytics***.
2. Clicar el botó ***Create a new query***.

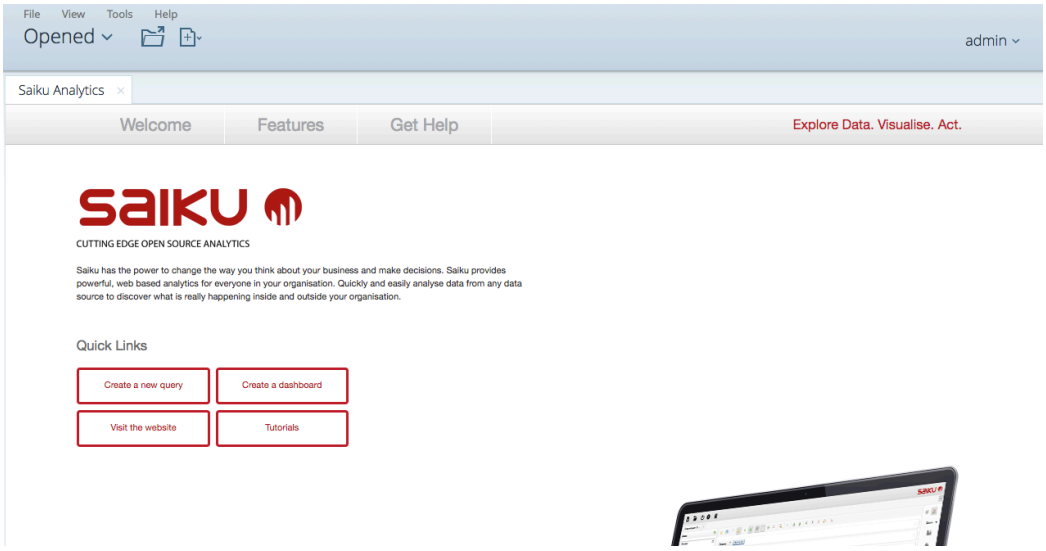



**Fig 23.** Menú d'opcions de *Saiku Analytics*

3. La creació de *querys* és molt senzilla gràcies a la interfície intuïtiva. Des del menú de l'esquerra podem escollir el cub que es vol analitzar entre tots els disponibles. Des de ***Medidas*** disposem de tots els indicadors del cub escollit. També disposem de les dimensions que formen part del cub seleccionat en l'apartat ***Dimensiones.*** A la barra superior disposem del menú que ens permetrà: navegar per directoris, desar, blanquejar la *query* actual, editar, executar la *query* de forma manual o automàtica, ocultar pares, no mostrar buits, intercanviar eixos, mostrar codi *MDX,* fer zoom, entre altres. *Saiku* també ens permet exportar els informes a varis formats com: Excel, CSV, o PDF. Per últim, al menú lateral disposem dels botons de canvi de visió de taules o de gràfics i els diferents models de gràfics disponibles.

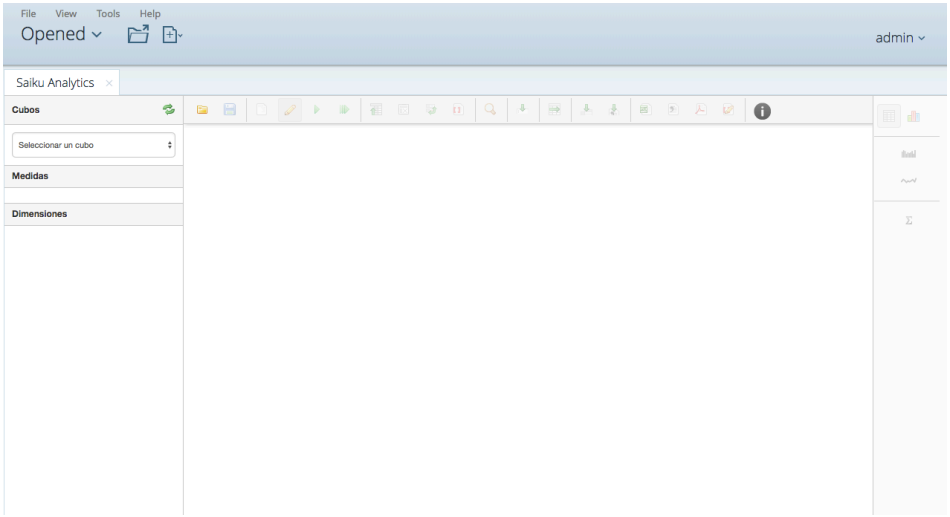



**Fig 24.** Interfície *Saiku Analytics*

4. Com a exemple s'ha fet una *query* del cub *Ventes* a partir dels filtres de les dimensions del cub. Com a mesura s'ha escollit *Ventes*, a les columnes s'ha seleccionat la *Família* i *Producte* de la *DimensioProductes* i la *Tipus* de la *DimensioPunts*, a les files s'ha seleccionat *Trimestre* de la *DimensioTemps*.

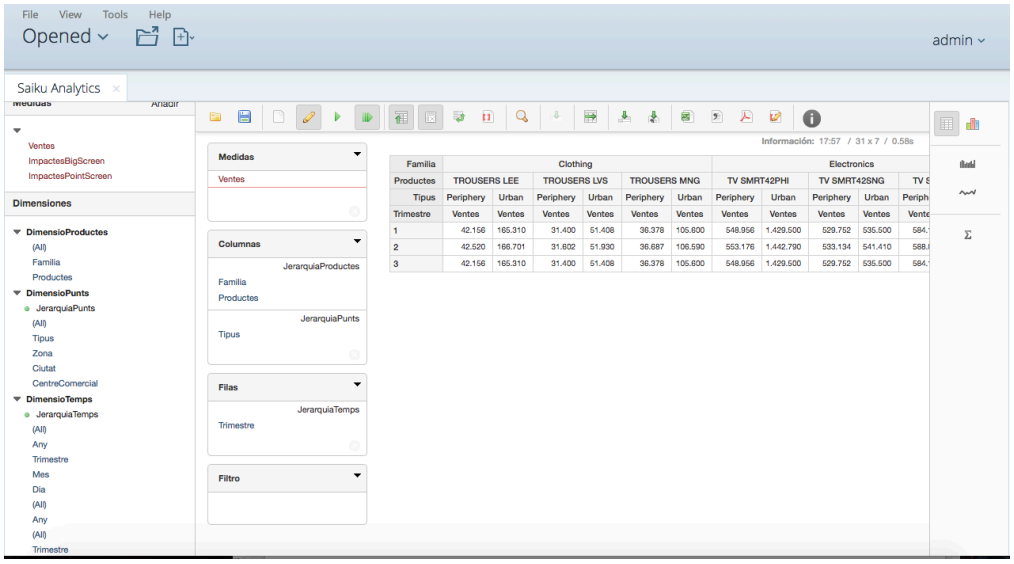



**Fig 25.** Query del cub de Ventes

## 4.4 Generació dels informes amb *Saiku Analytics*

Un cop descrits els passos per realitzar *querys* als cubs OLAP amb *Saiku Analytics* a l'apartat anterior, procedirem a continuació a crear les taules i gràfics que ens permetin donar resposta a les preguntes inicials del projecte.

Cadascun dels punt següents es centre en cadascuna de les preguntes plantejades al projecte i ens permetran argumentar les respostes i conclusions al final del projecte.

### 4.4.1 Informe d'efectivitat dels impactes visuals

Per analitzar l'efectivitat dels impactes visuals s'ha optat per comparar les ventes amb els impactes *BigScreen* i *PointScreen*. Els resultats es filtraran per l'indicador *Ventes, BigScreen* i *PointScreen* del cub de *Ventes*, segons el nivell *CentreComercial* de la jerarquia *Punts* de la dimensió *Punts*, per tal de reportar els resultats individuals per centre comercial. En aquest anàlisis, amb l'objectiu d'obtenir la quantitat de ventes totals segons els impactes visuals, no s'ha afegit la dimensió temps.

Les següents figures mostren les ventes i els impactes visuals per centre comercial:

| CentreComercial | Ventes | ImpactesPointScreen | ImpactesBigScreen | TotalImpactes |
|---|---|---|---|---|
| La Maquinista | 4.668.707 | 28.140 | 28.154 | 56.294 |
| Mercat del Pla | 4.165.409 | 28.095 | 28.267 | 56.362 |
| Gran Jonquera Outlet & Shopping | 3.609.643 | 28.178 | 28.227 | 56.405 |
| Girocentre | 3.598.122 | 28.198 | 28.208 | 56.406 |
| La Fira | 2.442.996 | 28.137 | 28.117 | 56.254 |
| Les Gabarres | 2.999.178 | 28.035 | 28.101 | 56.136 |

**Fig 26.** Nombre de ventes en relació als impactes distribuïts per centres

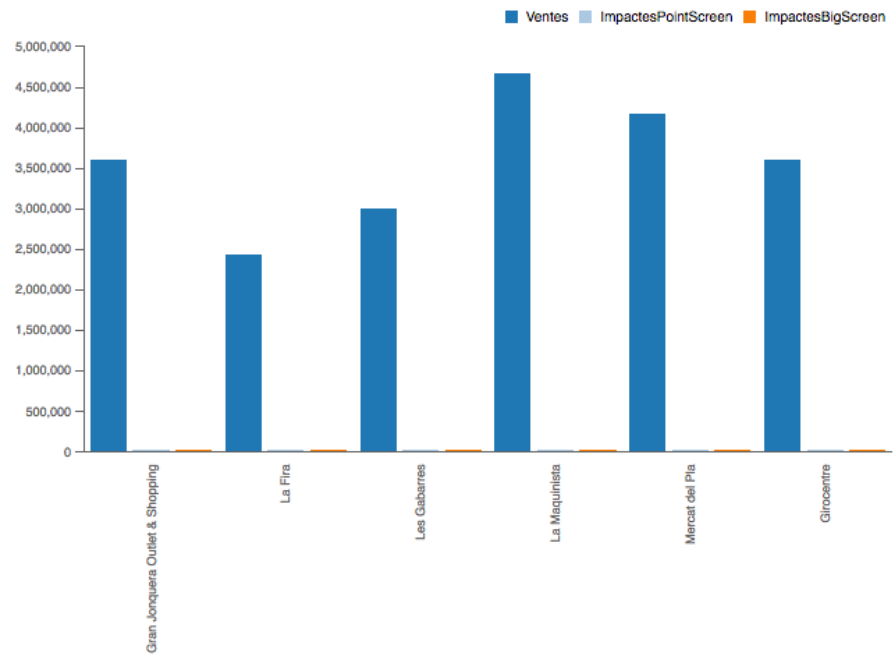



**Fig 27.** Distribució de les vendes i els impactes visuals de cada centre comercial

Es pot observar amb les figures anterior com a nivells generals els impactes visuals no influeixen directament a les ventes. Els sis centres comercials tenen, més o menys, els mateixos impactes visuals i en canvi hi ha diferències notables en les ventes. El centre comercial amb més ventes l'any 2016 és La Maquinista, en canvi, el centre amb més impactes visuals *PointScreen* és Girocentre. El centre que ha generat més impactes visuals *BigScreen* és Mercat del Pla que ocupa la segona posició en quantitat de ventes.

La observació anterior ens fa pensar que les visites a cada centre tinguin una clara importància en la influència dels impactes visuals sobre les ventes. És per aquest motiu que analitzarem les visites per centre comercial per veure si es confirmen les sospites.

| CentreComercial | Visites |
|---|---|
| Gran Jonquera Outlet & Shopping | 2.589.431 |
| La Fira | 2.566.992 |
| Les Gabarres | 4.430.682 |
| La Maquinista | 4.406.866 |
| Mercat del Pla | 2.291.391 |
| Girocentre | 2.609.648 |

**Fig 28.** Nombre de visites per centres comercials

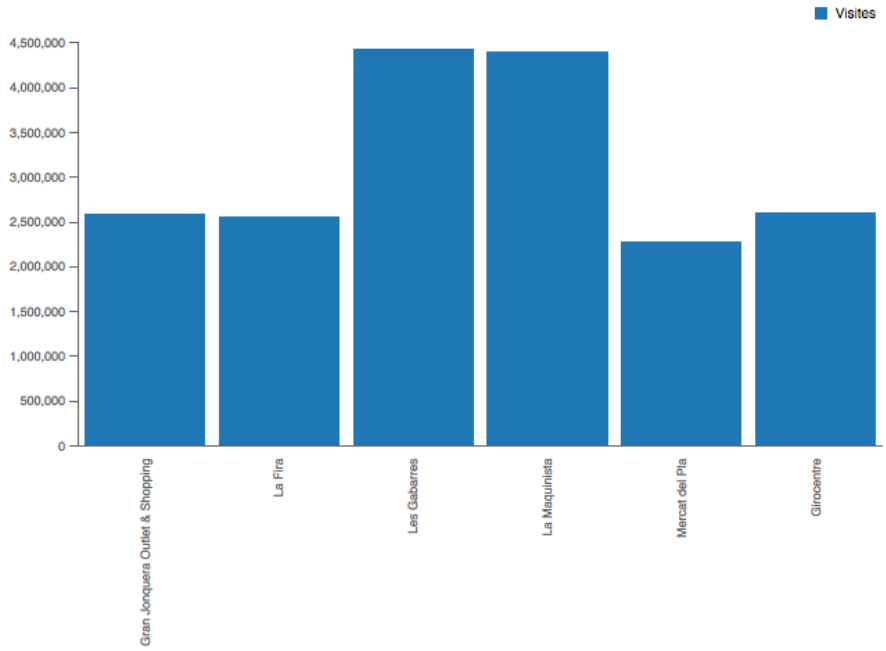



**Fig 29.** Distribució de visites per centre comercial

Com es pot observar amb les dues figures anteriors és evident que cal analitzar les visites als centres comercials per poder deduir si els impactes visuals influeixen directament o no en les ventes.

Observem que Les Gabarres és el centre comercial amb més visites l'any 2016, però en canvi, és el centre comercial amb el valor mínim d'impactes, fet que el converteix amb un dels centres amb una quantitat de ventes més baix. També podem observar com el centre comercial amb menys visites, Mercat del

Pla, és el segon posicionat en el nombre de ventes ja que és un dels centres amb una quantitat d'impactes visuals més alta.

### 4.4.2 Informe d'efectivitat dels impactes visuals segons les zones geogràfiques

En aquest informe analitzarem l'efectivitat dels impactes visuals segons les zones geogràfiques. Es compararà les ventes amb els impactes visuals segons la zona geogràfica en la que es troba el centre comercial, nivell *Zona* de la *DimensioPunts*.

Les següents figures mostren les ventes i impactes per zona geogràfica:

| Zona | Ventes | ImpactesPointScreen | ImpactesBigScreen | TotalImpactes |
|---|---|---|---|---|
| CENTER | 8.834.116 | 56.235 | 56.421 | 112.656 |
| NORTH | 7.207.765 | 56.376 | 56.435 | 112.811 |
| SOUTH | 5.442.174 | 56.172 | 56.218 | 112.390 |

**Fig 30.** Nombre de ventes i impactes segons zona geogràfica

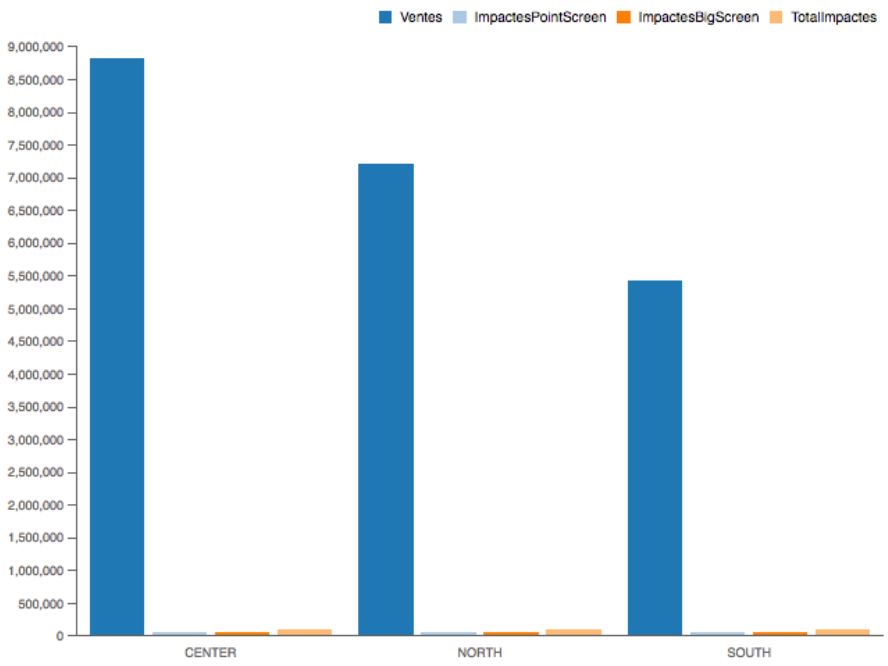



**Fig 31.** Distribució de ventes i impactes segons zona geogràfica

Com podem observar a les figures anteriors la zona amb més ventes és el Centre, i la zona amb menys ventes el Sud. Pel que fa als impactes visuals la zona que menys n'ha mostrat ha sigut, igual que en les ventes, el Sud.

Segons aquestes premisses podríem dir que els impactes visuals afecten en les ventes segons la zona geogràfica, ja que la zona amb menys impactes és la zona que ha tingut menys ventes. Però, si observem les altres dues zones veiem que la zona Nord té una mica més d'impactes visuals que la zona Centre, en canvi, el Nord està per sota de les ventes del Centre.

Després d'aquest segon anàlisis es pot continuar deduint que els impactes visuals no afecten, almenys de forma considerable, a les ventes dels centres comercials.

Tal i com hem fet a l'informe anterior analitzarem les visites per zona geogràfica per poder afirmar si els impactes influeixen a les ventes.

| Zona | Visites |
|---|---|
| CENTER | 6.698.257 |
| NORTH | 5.199.079 |
| SOUTH | 6.997.674 |

**Fig 32.** Nombre de visites per zona geogràfica

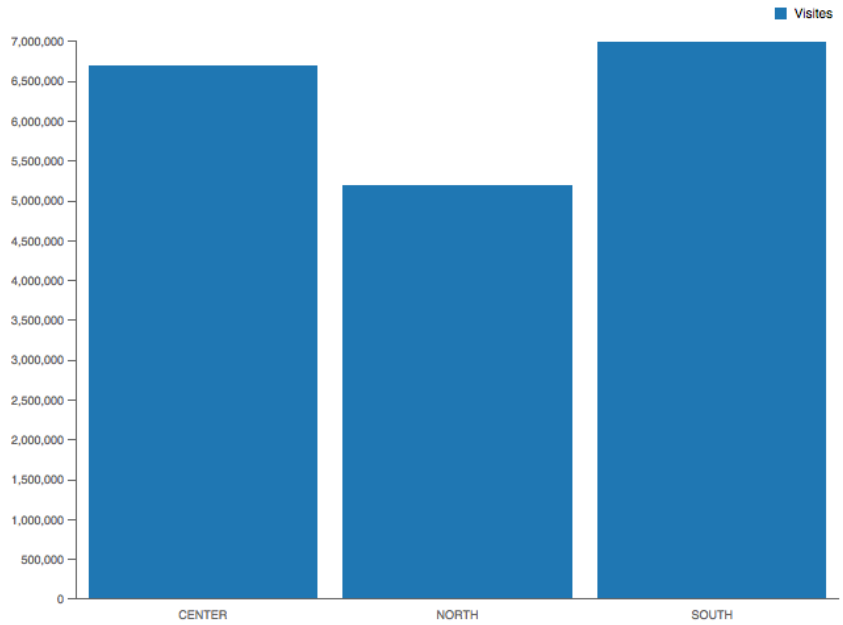



**Fig 33.** Distribució de vistes per zona geogràfica

Observem que la zona geogràfica amb més visites és el Sud però en canvi és la zona amb menys quantitat d'impactes i en conseqüència de ventes. En canvi, el Nord és la zona amb menys visites i en canvi té el nombre més alt d'impactes, fet que és reflecteix en la quantitat de ventes, bastant superior a les del Sud.

Per tant, podem afirmar que segons la zona geogràfica la publicitat pot ser més efectiva o no ja que una zona que ha tingut més visites que un altre però que en canvi ha tingut un nombre semblant d'impactes, té un valor total de ventes inferior.

### 4.4.3 Informe d'efectivitat dels impactes visuals segons la família del producte

En aquest informe s'ha optat per comparar, un altre vegada, les ventes amb els impactes visuals agrupats per família dels productes.

| Familia | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
|---|---|---|---|---|
| Clothing | 1.173.039 | 33.808 | 33.781 | 67.589 |
| Electronics | 15.606.459 | 33.820 | 33.609 | 67.429 |
| Food | 752.731 | 33.639 | 33.820 | 67.459 |
| Home | 769.029 | 33.951 | 33.818 | 67.769 |
| Sports | 3.182.797 | 33.856 | 33.755 | 67.611 |

**Fig 34.** Nombre de ventes i impactes visuals per família de producte

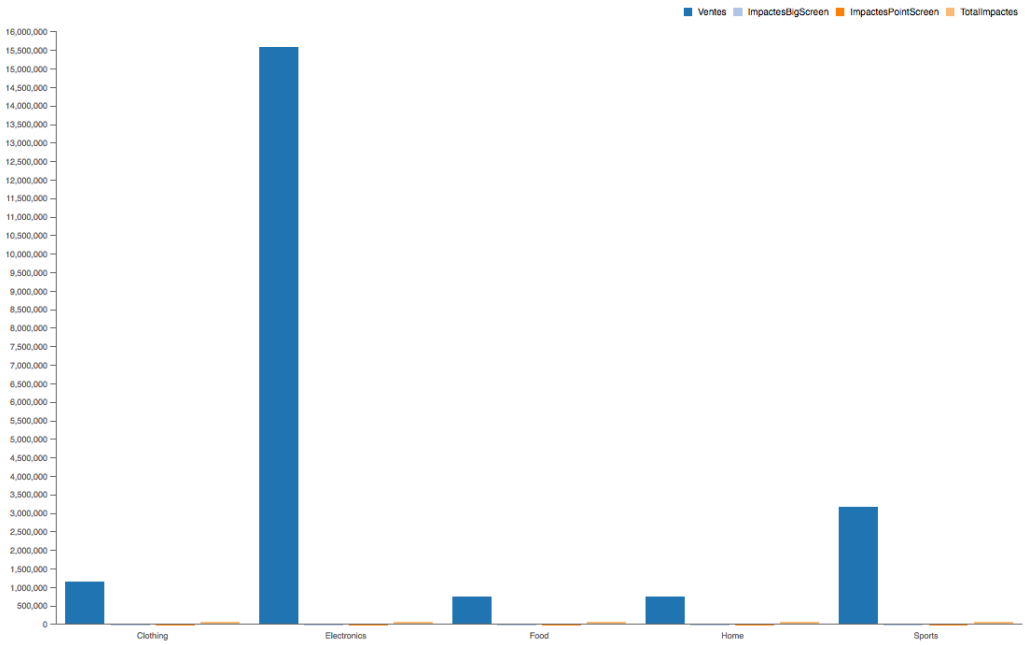



**Fig 35.** Distribució de ventes i impactes visuals per família de producte

Observem que la família de productes electrònics és amb diferència la que més quantitat de ventes ha obtingut l'any 2016, en canvi, el valor total d'impactes visuals és gairebé el mateix en totes les famílies.

Podríem afirmar doncs, que els impactes visuals no tenen efecte sobre les ventes segons la família del producte. Tot i així, hem de tenir en compte que no tots els articles tenen el mateix preu ni el mateix marge de benefici. No ens podem guiar per la quantitat (en euros) de ventes d'una família o una altre ja que un producte d'electrònica sempre serà molt més car que un producte de menjar i per tant, les ventes totals dels articles més cars sempre seran molt superiors a la resta.

Seguidament analitzarem si dins dels productes d'una mateixa família, podem trobar una relació entre els articles impactats i el resultat obtingut. Per això, redefinirem les figures anteriors afegint el nivell producte de la DimensioProductes.

| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
|---|---|---|---|---|---|
| Clothing | TROUSERS LEE | 441.843 | 13.715 | 13.687 | 27.402 |
| | TROUSERS LVS | 344.258 | 9.113 | 9.038 | 18.151 |
| | TROUSERS MNG | 386.938 | 10.980 | 11.056 | 22.036 |
| Electronics | TV SMRT42PHI | 5.132.042 | 11.011 | 10.921 | 21.932 |
| | TV SMRT42SNG | 4.549.278 | 9.036 | 8.953 | 17.989 |
| | TV SMRT42SNY | 5.925.139 | 13.773 | 13.735 | 27.508 |
| Food | PIZZA CUSN | 291.084 | 13.717 | 13.746 | 27.463 |
| | PIZZA DROE | 247.546 | 10.893 | 11.003 | 21.896 |
| | PIZZA TDRL | 214.101 | 9.029 | 9.071 | 18.100 |
| Home | FRAME 18X10 CH | 253.657 | 11.061 | 10.993 | 22.054 |
| | FRAME 18X10 CK | 240.168 | 9.092 | 9.076 | 18.168 |
| | FRAME 18X10 MD | 275.204 | 13.798 | 13.749 | 27.547 |
| Sports | SHOES ACS | 1.253.045 | 13.786 | 13.677 | 27.463 |
| | SHOES ADS | 1.038.346 | 10.954 | 11.041 | 21.995 |
| | SHOES NKE | 891.406 | 9.116 | 9.037 | 18.153 |

**Fig 36.** Nombre de ventes i impactes per família i producte

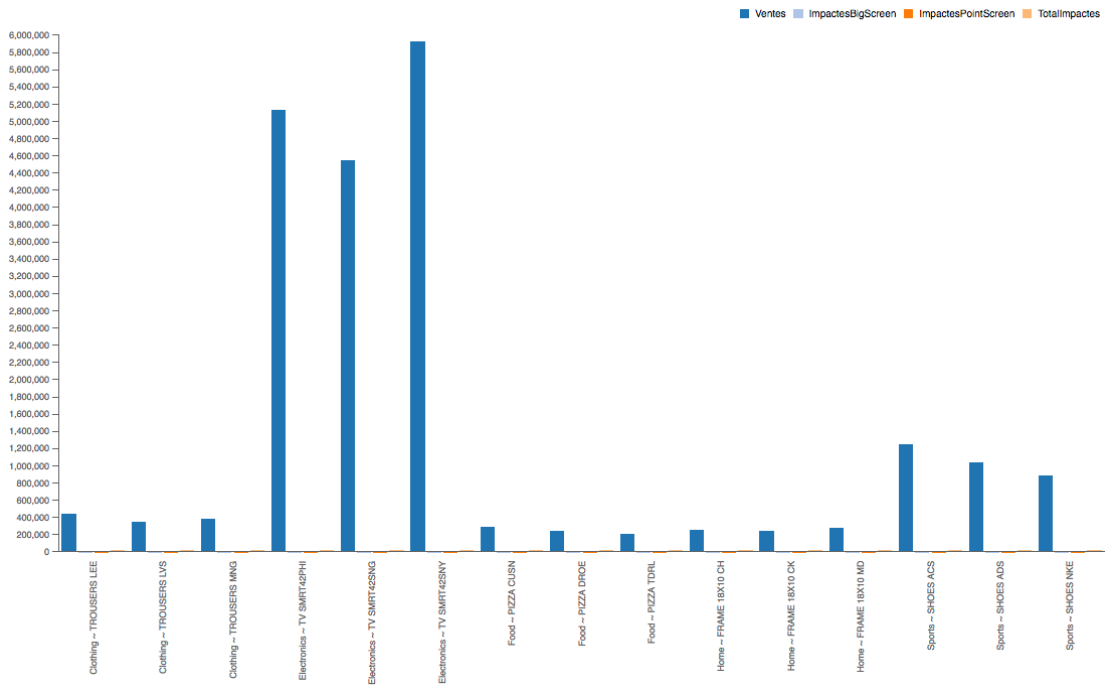



**Fig 37.** Distribució de ventes i impactes per família i producte

Com podem observar en les noves figures, dins de cada família l'ordre en nombre d'impactes és el mateix ordre en nombre de ventes. Per tant, podem afirmar que com més impactes té un article més es vendrà.

### 4.4.4 Informe d'efectivitat dels impactes visuals segons el moment temporal

Per analitzar l'efectivitat dels impactes visuals s'ha optat per comparar les ventes amb els impactes *BigScreen* i *PointScreen*. Els resultats es filtraran per l'indicador *Ventes, BigScreen* i *PointScreen* del cub de *Ventes*, segons el nivell *Quatrimestre* de la jerarquia *Temps* de la dimensió *DimensióTemps*, per tal de reportar els resultats individuals per quadrimestre.

| Quadrimestre | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
|---|---|---|---|---|
| 1 | 8.070.498 | 67.411 | 67.223 | 134.634 |
| 2 | 8.142.131 | 68.530 | 68.335 | 136.865 |
| 3 | 8.070.498 | 67.984 | 67.794 | 135.778 |

**Fig 38.** Nombre de ventes i impactes visuals per quadrimestre.

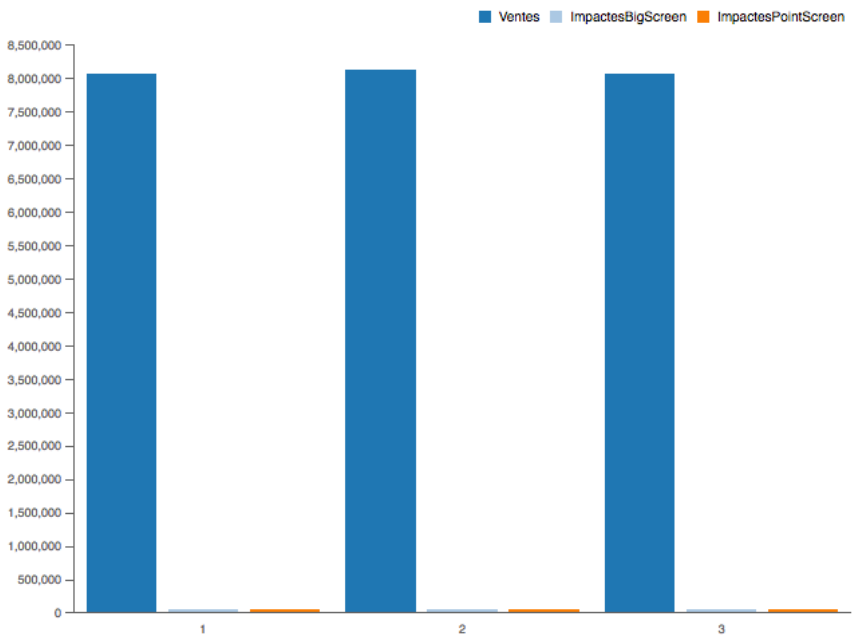



**Fig 39.** Distribució de ventes i impactes visuals per quadrimestre

Observant les figures anteriors veiem que el segon quadrimestre és el que té el valor màxim de ventes i d'impactes visuals. En canvi, els altres dos quadrimestres estan igualats en ventes però el tercer quadrimestre té un nombre d'impactes visuals lleugerament superior als del primer quadrimestre.

Per tal de poder analitzar amb més detall si el moment temporal dels impactes visuals afecta directament a les ventes, analitzarem les visites distribuïdes per quadrimestres.

| Quadrimestre | Visites |
|---|---|
| 1 | 6.291.995 |
| 2 | 7.631.922 |
| 3 | 4.875.030 |

**Fig 40.** Nombre de visites per quadrimestre

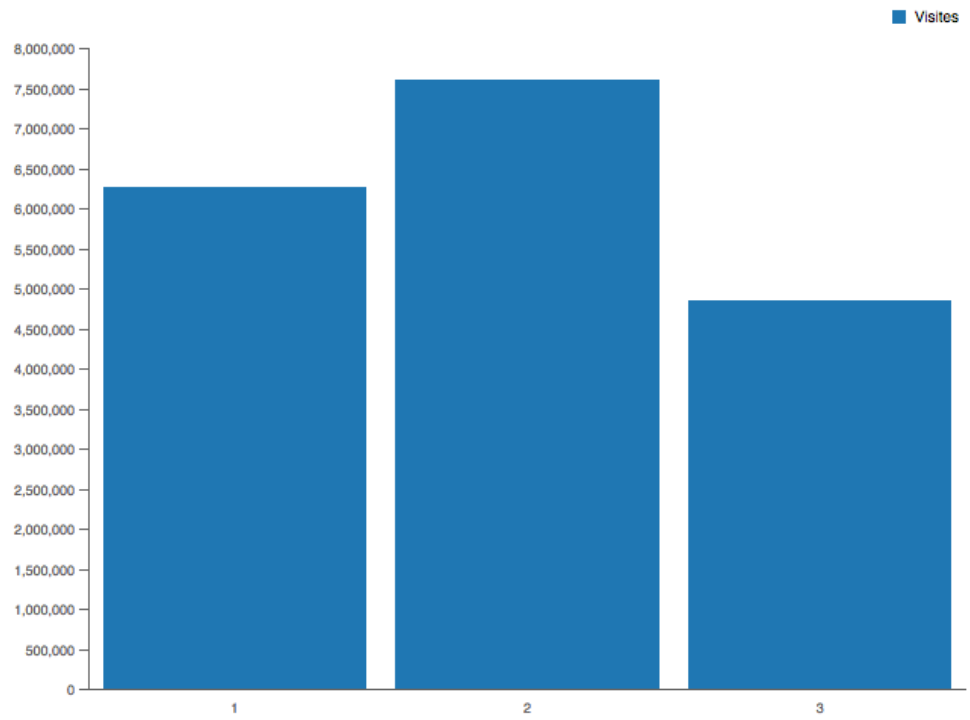



**Fig 41.** Distribució del nombre de visites per quadrimestre

Les figures anteriors ens indiquen que el quadrimestre amb menys visites és el tercer però, al tenir un nombre d'impactes superior al del primer quadrimestre, ha aconseguit igualar les ventes d'aquest.

Per tant, podem afirmar que en moments temporals on les visites disminueixen considerablement respecte la resta de l'any, els impactes visuals són efectius i permeten que els centres comercials augmentin la quantitat de ventes.

### 4.4.5 Informe d'efectivitat dels impactes visuals segons la zona i l'article o família

En aquest informe analitzarem si existeix algun article o família que segons la zona la seva publicitat tingui un major impacte sobre els ventes. Es compararà les ventes amb els impactes visuals segons la zona geogràfica en la que es troba el centre comercial, nivell *Zona* de la *DimensioPunts*.

| Zona | | CENTER | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 201.577 | 4.544 | 4.560 | 9.104 |
| | TROUSERS LVS | 152.406 | 3.037 | 3.016 | 6.053 |
| | TROUSERS MNG | 174.151 | 3.672 | 3.738 | 7.410 |
| Electronics | TV SMRT42PHI | 2.533.950 | 3.682 | 3.636 | 7.318 |
| | TV SMRT42SNG | 2.121.120 | 3.002 | 2.970 | 5.972 |
| | TV SMRT42SNY | 3.108.840 | 4.631 | 4.589 | 9.220 |
| Food | PIZZA CUSN | 34.229 | 4.605 | 4.587 | 9.192 |
| | PIZZA DROE | 34.165 | 3.600 | 3.664 | 7.264 |
| | PIZZA TDRL | 33.987 | 3.007 | 3.041 | 6.048 |
| Home | FRAME 18X10 CH | 78.152 | 3.673 | 3.664 | 7.337 |
| | FRAME 18X10 CK | 68.189 | 3.078 | 3.005 | 6.083 |
| | FRAME 18X10 MD | 93.676 | 4.586 | 4.569 | 9.155 |
| Sports | SHOES ACS | 66.218 | 4.607 | 4.529 | 9.136 |
| | SHOES ADS | 67.342 | 3.644 | 3.675 | 7.319 |
| | SHOES NKE | 66.114 | 3.053 | 2.992 | 6.045 |

| Zona | | NORTH | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 184.280 | 4.583 | 4.545 | 9.128 |
| | TROUSERS LVS | 136.790 | 3.044 | 3.040 | 6.084 |
| | TROUSERS MNG | 157.385 | 3.644 | 3.643 | 7.287 |
| Electronics | TV SMRT42PHI | 1.682.160 | 3.698 | 3.702 | 7.400 |
| | TV SMRT42SNG | 1.521.169 | 3.046 | 3.035 | 6.081 |
| | TV SMRT42SNY | 1.895.015 | 4.565 | 4.558 | 9.123 |
| Food | PIZZA CUSN | 97.916 | 4.554 | 4.569 | 9.123 |
| | PIZZA DROE | 82.105 | 3.694 | 3.648 | 7.342 |
| | PIZZA TDRL | 72.269 | 3.005 | 3.026 | 6.031 |
| Home | FRAME 18X10 CH | 91.477 | 3.669 | 3.675 | 7.344 |
| | FRAME 18X10 CK | 88.717 | 3.007 | 3.033 | 6.040 |
| | FRAME 18X10 MD | 95.117 | 4.632 | 4.599 | 9.231 |
| Sports | SHOES ACS | 437.937 | 4.569 | 4.592 | 9.161 |
| | SHOES ADS | 356.466 | 3.708 | 3.689 | 7.397 |
| | SHOES NKE | 308.962 | 3.017 | 3.022 | 6.039 |

| Zona | | SOUTH | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 55.986 | 4.588 | 4.582 | 9.170 |
| | TROUSERS LVS | 55.062 | 3.032 | 2.982 | 6.014 |
| | TROUSERS MNG | 55.402 | 3.664 | 3.675 | 7.339 |
| Electronics | TV SMRT42PHI | 915.932 | 3.631 | 3.583 | 7.214 |
| | TV SMRT42SNG | 906.989 | 2.988 | 2.948 | 5.936 |
| | TV SMRT42SNY | 921.284 | 4.577 | 4.588 | 9.165 |
| Food | PIZZA CUSN | 158.939 | 4.558 | 4.590 | 9.148 |
| | PIZZA DROE | 131.276 | 3.599 | 3.691 | 7.290 |
| | PIZZA TDRL | 107.845 | 3.017 | 3.004 | 6.021 |
| Home | FRAME 18X10 CH | 84.028 | 3.719 | 3.654 | 7.373 |
| | FRAME 18X10 CK | 83.262 | 3.007 | 3.038 | 6.045 |
| | FRAME 18X10 MD | 86.411 | 4.580 | 4.581 | 9.161 |
| Sports | SHOES ACS | 748.890 | 4.610 | 4.556 | 9.166 |
| | SHOES ADS | 614.538 | 3.602 | 3.677 | 7.279 |
| | SHOES NKE | 516.330 | 3.046 | 3.023 | 6.069 |

**Fig 42.** Nombre de ventes i impactes per família i producte i zones

Tal i com hem pogut observar a la secció 4.4.3 no podem comparar els impactes de la publicitat entre famílies ja que el valor dels articles és molt diferent i ens faltaria informació addicional per poder-les comparar. El que si podem comparar són els articles d'una mateixa família.

Com podem observar a les figures anteriors, excepte en els productes d'esports de la zona centre, la resta dins d'una mateixa família sempre té un valor més elevat de ventes l'article que disposa de més impactes visuals.

Per poder analitzar si la publicitat és més efectiva en un producte o un altre segons la zona geogràfica haurem d'analitzar també les visites als centres comercials per zona.

| Zona | Visites |
|---|---|
| CENTER | 6.698.257 |
| NORTH | 5.199.079 |
| SOUTH | 6.997.674 |

Si ens fixem en les visites per zona, els impactes visuals i les ventes per zona, podem afirmar que la publicitat és més efectiva en un producte segons la zona geogràfica. Això és veu perfectament si ens fixem, per exemple, en els

productes de roba. La zona sud disposa de més visites que la zona centre i a la vegada tenen el mateix nombre d'impactes visuals en els tres productes, en canvi, la zona centre tot i tenir menys visites, supera de llarg les ventes de la zona sud. Per tant, la publicitat de productes de roba en la zona centre és més efectiva que en la zona sud. Això també passa amb la família de producte electrònics. En canvi, en el menjar i els esports, la publicitat té més efecte en la zona nord i sud que en el centre.

Finalment, podem dir que on la publicitat té un efecte més gran és en els productes de casa de la zona nord, ja que tot i que sigui la zona amb menys visites, és la zona on amb el mateix nombre d'impactes, la quantitat de ventes és més alta.

### 4.4.6 Informe d'efectivitat dels impactes visuals segons el moment temporal i l'article o família

En aquest informe analitzarem si existeix algun article o família que segons el moment temporal la seva publicitat tingui un major impacte sobre els ventes. Es comparará les ventes amb els impactes visuals segons la zona geogràfica en la que es troba el centre comercial, nivell *Quadrmiestre* de la *DimensioTemps*.

| JerarquiaTemps - Quadrimestre | | 1 | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 207.466 | 7.660 | 7.689 | 15.349 |
| | TROUSERS LVS | 82.808 | 1.530 | 1.479 | 3.009 |
| | TROUSERS MNG | 141.978 | 4.284 | 4.412 | 8.696 |
| Electronics | TV SMRT42PHI | 1.978.456 | 4.342 | 4.361 | 8.703 |
| | TV SMRT42SNG | 1.065.252 | 1.455 | 1.407 | 2.862 |
| | TV SMRT42SNY | 2.847.748 | 7.638 | 7.610 | 15.248 |
| Food | PIZZA CUSN | 143.048 | 7.683 | 7.686 | 15.369 |
| | PIZZA DROE | 93.036 | 4.294 | 4.374 | 8.668 |
| | PIZZA TDRL | 46.890 | 1.537 | 1.408 | 2.945 |
| Home | FRAME 18X10 CH | 89.668 | 4.427 | 4.319 | 8.746 |
| | FRAME 18X10 CK | 68.010 | 1.466 | 1.465 | 2.931 |
| | FRAME 18X10 MD | 112.890 | 7.704 | 7.623 | 15.327 |
| Sports | SHOES ACS | 631.068 | 7.658 | 7.660 | 15.318 |
| | SHOES ADS | 390.270 | 4.305 | 4.279 | 8.584 |
| | SHOES NKE | 171.910 | 1.428 | 1.451 | 2.879 |

| JerarquiaTemps - Quadrimestre | | 2 | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 209.221 | 7.787 | 7.815 | 15.602 |
| | TROUSERS LVS | 83.532 | 1.556 | 1.504 | 3.060 |
| | TROUSERS MNG | 143.277 | 4.356 | 4.481 | 8.837 |
| Electronics | TV SMRT42PHI | 1.995.966 | 4.407 | 4.434 | 8.841 |
| | TV SMRT42SNG | 1.074.544 | 1.480 | 1.435 | 2.915 |
| | TV SMRT42SNY | 2.872.462 | 7.763 | 7.739 | 15.502 |
| Food | PIZZA CUSN | 144.476 | 7.805 | 7.817 | 15.622 |
| | PIZZA DROE | 94.008 | 4.370 | 4.456 | 8.826 |
| | PIZZA TDRL | 47.299 | 1.563 | 1.431 | 2.994 |
| Home | FRAME 18X10 CH | 90.472 | 4.508 | 4.386 | 8.894 |
| | FRAME 18X10 CK | 68.631 | 1.487 | 1.487 | 2.974 |
| | FRAME 18X10 MD | 113.891 | 7.835 | 7.742 | 15.577 |
| Sports | SHOES ACS | 637.522 | 7.786 | 7.793 | 15.579 |
| | SHOES ADS | 393.345 | 4.373 | 4.343 | 8.716 |
| | SHOES NKE | 173.485 | 1.454 | 1.472 | 2.926 |

| JerarquiaTemps - Quadrimestre | | 3 | | | |
|---|---|---|---|---|---|
| Familia | Productes | Ventes | ImpactesBigScreen | ImpactesPointScreen | TotalImpactes |
| Clothing | TROUSERS LEE | 207.466 | 7.726 | 7.754 | 15.480 |
| | TROUSERS LVS | 82.808 | 1.544 | 1.492 | 3.036 |
| | TROUSERS MNG | 141.978 | 4.320 | 4.446 | 8.766 |
| Electronics | TV SMRT42PHI | 1.978.456 | 4.378 | 4.396 | 8.774 |
| | TV SMRT42SNG | 1.065.252 | 1.468 | 1.422 | 2.890 |
| | TV SMRT42SNY | 2.847.748 | 7.702 | 7.674 | 15.376 |
| Food | PIZZA CUSN | 143.048 | 7.746 | 7.754 | 15.500 |
| | PIZZA DROE | 93.036 | 4.332 | 4.420 | 8.752 |
| | PIZZA TDRL | 46.890 | 1.550 | 1.418 | 2.968 |
| Home | FRAME 18X10 CH | 89.668 | 4.468 | 4.356 | 8.824 |
| | FRAME 18X10 CK | 68.010 | 1.478 | 1.478 | 2.956 |
| | FRAME 18X10 MD | 112.890 | 7.770 | 7.684 | 15.454 |
| Sports | SHOES ACS | 631.068 | 7.720 | 7.726 | 15.446 |
| | SHOES ADS | 390.270 | 4.342 | 4.310 | 8.652 |
| | SHOES NKE | 171.910 | 1.440 | 1.464 | 2.904 |

**Fig 43.** Nombre de ventes i impactes visuals per família, producte i quadrimestre

Per poder analitzar si la publicitat és més efectiva en un producte o un altre segons el moment temporal haurem d'analitzar també les visites als centres comercials per quadrimestre.

| Quadrimestre | Visites |
|---|---|
| 1 | 6.291.995 |
| 2 | 7.631.922 |
| 3 | 4.875.030 |

Si ens fixem en les visites, els impactes visuals i les ventes per quadrimestre, podem afirmar que la publicitat és més efectiva en un producte segons el

moment temporal. Això és veu perfectament si ens fixem, per exemple, en els productes de roba. El primer quadrimestre disposa de més visites que el tercer quadrimestre i a la vegada té el mateix nombre d'impactes visuals que els altres quadrimestres, en canvi, és superat en quantitat de ventes per el tercer quadrimestre tot i que aquest tingui menys visites. Per tant, la publicitat de productes de roba en el tercer quadrimestre és més efectiva que en el primer. En la resta de famílies la tendència és la mateixa i el segon quadrimestre és el que queda en pitjor posició ja que tot i tenir més visites i el mateix nombre d'impactes no aconsegueix superar per gaire en ventes als altres dos quadrimestres.

## 4.5 Resposta a les preguntes plantejades

Un cop realitzats els informes anteriors procedirem a donar resposta a les preguntes plantejades al projecte.

**Pregunta 1.** En general la publicitat basada en impactes visuals, és efectiva?

Si, tal i com hem vist als informes anteriors en general la publicitat basades en impactes visuals és efectiva. En l'informe 4.4.1 hem vist com, en general, els centres comercials que amb el mateix nivell de visites tenen un valor més alt d'impactes visuals, també tenen una quantitat més elevada de ventes. També hem vist que per poder analitzar l'efecte de la publicitat cal tenir molt en compte el nombre de visites a cada centre per poder donar una resposta realista.

**Pregunta 2.** Hi ha determinades zones geogràfiques on els impactes són més efectius?

Si, tal i com hem pogut veure al informe 4.4.2, la zona sud té menys visites que la zona centre i un nombre d'impactes molt semblant, en canvi, el centre ha obtingut un nombre molt superior en ventes. Per tant, podem afirmar que els impactes són més efectius en la zona centre que en la zona sud. El mateix passa amb la zona nord que tot i tenir més impactes visuals disposa de moltes

menys visites que la zona sud, no obstant, les ventes són molt superiors a les del sud.

**Pregunta 3.** Segons la família de producte, existeix alguna relació entre els articles impactats i el resultat obtingut?

Tal i com hem pogut comprovar en l'informe 4.4.3 no podem afirmar si existeix una relació entre els articles impactats i el resultat obtingut segons la família del producte. No podem comparar les ventes de dues famílies diferents amb la informació de la que disposem. Entre famílies el preu dels productes són molt diferents i per tant, no ens podem fiar de les dades que tenim per extreure'n conclusions.

Una manera per poder analitzar aquesta pregunta seria disposar del benefici net que reporta al centre comercial les ventes de cada família d'articles i el cost que té la publicitat de cada família. D'aquesta manera podríem saber quina de les famílies reporta més beneficis nets al centre i permetria als comerciants saber en quins productes han d'invertir més o menys publicitat per extreure el màxim rendiment de la publicitat.

Tot i així, si que podem analitzar un producte respecte els altres productes de la mateixa família. Al final de l'informe 4.4.3 tenim les gràfiques que ens demostren que el nombre d'impactes influeix directament a les ventes, ja que sempre obté més ventes el producte que té més impactes visuals.

**Pregunta 4.** Són més efectius els impactes segons el moment temporal en que es produeixen?

Si. Tal i com hem vist a l'informe 4.4.4 hi ha quadrimestres on les visites són considerablement inferiors als de la resta però amb un nombre semblant d'impactes visuals obtenen beneficis superiors en ventes.

Un exemple clar és el que passa amb el tercer trimestre, tot i tenir unes visites molt inferiors a la resta, i un nombre d'impactes visuals semblants a la resta de

quadrimestres, iguala en ventes al primer quadrimestre i és queda poc per sota del segon quadrimestre.

**Pregunta 5.** Existeix algun article o família que segons la zona, la seva publicitat té major impacte? I segons el moment en que es realitza?

Segons hem pogut veure en l'informe 4.4.5 hi ha famílies com els productes de roba que, per exemple, la seva publicitat reporta més ventes si es fan en un centre comercial situat en la zona centre que no pas en la zona sud. En altres famílies com el menjar i els esports la publicitat té major impacte en la zona nord i el centre. Per acabar, la publicitat té un efecte més gran és en els productes de casa de la zona nord, ja que tot i que sigui la zona amb menys visites, és la zona on amb el mateix nombre d'impactes, la quantitat de ventes és més alta.

Per altre banda, segons el moment en que es realitza l'impacte, la publicitat és més efectiva o no. Això ho hem analitzat en l'informe 4.4.6. Veiem com el primer quadrimestre tot i tenir més visites i els mateixos impactes té menys ventes que el tercer quadrimestre. És a dir, la publicitat de productes de roba en el tercer quadrimestre és més efectiva que en el primer. En la resta de famílies la tendència és la mateixa i el segon quadrimestre és el que queda en pitjor posició ja que tot i tenir més visites i el mateix nombre d'impactes no aconsegueix superar per gaire en ventes als altres dos quadrimestres.

# 5. Conclusions

Al llarg d'aquest TFM s'han anat extraient una sèrie de conclusions que venen a representar l'objectiu final del treball realitzat. Es pot concloure que:

- Les eines de *Business Intelligence* que hi ha actualment al mercat, donen un ampli ventall de possibilitats d'anàlisis de dades a les organitzacions que necessiten convertir aquestes dades en informació útil que els permeti prendre decisions positives per al seu negoci. Un gran avantatge d'aquestes eines és que les *Open Source* disponibles al mercat, poden adaptar-se a cada entorn de negoci i cobrir perfectament les necessitats de qualsevol empresa.

- Una de les parts més importants per qualsevol organització és dur a terme un procés de disseny i implementació del *Data Warehouse* adequat per emmagatzemar totes les dades dels seus sistemes d'informació d'una forma eficient i escalable. Si el *Data Warehouse* no està adequadament dissenyat els processos d'extracció d'informació posteriors seran molt costosos o fins i tot podrien no ajustar-se a la realitat.

- Els processos *ETL* és una de les parts més complicades durant el procés de disseny del *Data Warehouse* ja que les dades d'una companyia no sempre vindran tal i com les necessitem.

- Les estructures de dades OLAP permeten fer abstraccions precises i concretes de la informació emmagatzemada en els *Data Warehouse* més complexos, permeten a les companyies realitzar anàlisis profunds dels conjunts de dades.

- *Pentaho* consta d'un seguit d'eines BI *Open Source* i una extensa comunitat d'usuaris que ens ha permès assolir els objectius del projecte de forma fàcil i senzilla.

- Les dades d'una companyia normalment no venen donades de forma que es puguin manipular directament amb eines com *Pentaho*, sinó que han de modificades prèviament amb processos ETL per poder disposar d'elles i així poder-les exportar al *Data Warehouse* dissenyat adequadament.

Els objectius marcats a l'inici del projecte s'han assolit amb èxit. S'ha dissenyat un *Data Warehouse* que permet emmagatzemar la informació adquirida dels diferents orígens. A més a més, s'ha dissenyat de forma que és un sistema escalable en el que es podria afegir informació d'altres centres comercials sense haver de modificar l'estructura. Un alter objectiu assolit amb èxit és que el *Data Warehouse* està dissenyat per poder-se alimentar directament dels fitxers base facilitats, aplicant prèviament els scripts per alimentar els esquemes de bases de dades anteriors al definitiu *Data Warehouse*.

S'ha escollit també, després d'un anàlisis de les diferents opcions, una plataforma BI *Open Source* que ens permet explotar amb èxit la informació emmagatzemada.

Finalment, s'ha intentat crear un quadre de comandament amb *Pentaho CDE Dashboards* per representar els indicadors claus però ha sigut impossible. No he aconseguit que la opció OLAP *Chart Wizard* funciones ja que al intentar construir el gràfic el servidor de *Pentaho* es quedava carregant per un temps il·limitat i no em generava els gràfics. Després de no trobar una solució al problema buscant dins de la comunitat de *Pentaho* ni una alternativa a *Pentaho CDE Dashboards* he decidit no generar el *Dashboard*.

Cal dir que l'idea inicial era aplicar els processos ETL amb l'eina *Pentaho Data Integration*, però després de realitzar una investigació del seu funcionament, és va optar per utilitzar processos SQL per a la implementació final del *Data Warehouse*. El motiu principal és que és una eina molt complerta amb moltes possibilitats però que requereix d'un temps d'aprenentatge que hagués fet que la resta de fites s'enrederissin considerablement.

Finalment, a nivell personal el projecte m'ha semblat molt interessant del principi al final. Professionalment em dedico a dissenyar aplicacions web per a gestió empresarial i segur que els coneixements adquirits durant aquest treball em permetran poder ampliar les funcionalitats d'aquestes aplicacions. De la mateixa manera segur que en un futur podria utilitzar eines com *Pentaho* o similars per ajudar als meus clients amb la gestió de la informació.

# 6. Glossari

***Business Intelligence.*** Es denomina intel·ligència empresarial, intel·ligència de negoci o BI, al conjunt d'estratègies i aspectes rellevants enfocats a l'administració i creació de coneixement sobre el mitjà, a través de l'anàlisi de les dades existents a una organització o empresa.

***Dashboard.*** És una interfície gràfica centrada en l'usuari que integra dades d'acord amb els problemes, funcions principals o processos comercials crítics de l'empresa. Freqüentment estan dissenyats per tractar un únic problema de forma aïllada i desenvolupar des de simples informes en línia fins a una complexa representació visual d'indicadors clau. Un dashboard de distribució pot incloure entre 20 i 25 indicadors diferents per establir la eficiència i la qualitat dels espais d'emmagatzemament, mesurats diàriament i representats en taules amb figures i gràfics, diagrames, agulles o rellotges.

***Data Warehouse.*** En el context de la informàtica, un magatzem de dades és una col·lecció de dades orientada a un determinat àmbit (empresa, organització, etc.), integrat, no volàtil i variable en el temps, que ajuda a la presa de decisions en l'entitat a la que s'utilitza.

***ETL.*** Extract, transform and Load (extreure, transformar i carregar) és el procés que permet a les organitzacions moure dades des de múltiples fonts, reformatetjar-los i netejar-los, i carregar-los a una altra base de dades, data mart, o Data Warehouse per analitzar, o en un altre sistema operacional per recolzar un procés de negoci.

***Framework.*** En el desenvolupament de programari, és una estructura conceptual i tecnològica de suport definit, normalment amb artefactes o mòduls de programari concrets, que pot servir de base per l'organització i desenvolupament de programari. Típicament, pot incloure suport de programes, biblioteques i un llenguatge interpretat, entre altres eines, per ajudar a desenvolupar i unir els diferents components d'un projecte.

***MDX (MultiDimesional eXpressions).*** És un llenguatge de consulta per bases de dades multidimensionals sobre cubs OLAP, s'utilitza en Business Intelligence per generar reports per la presa de decisions basades en dades històriques , amb la possibilitat de canviar l'estructura o rotació del cub.

***OLAP. (On-Line Analytical Processing).*** És una solució utilitzada en el camp de l'anomenada intel·ligència de negocis (Business Intelligence) que té com objectiu agilitzar la consulta de grans quantitats de dades. Utilitza estructures multidimensionals (o cubs OLAP) que contenen dades resumides de grans bases de dades o sistemes transaccionals (OLTP). S'utilitza en informes de negocis de vendes, marketing, informes de direcció, mineria de dades i àrees semblants.

***OLTP (OnLine Transaction Processing).*** És un tipus de processament que facilita i administra aplicacions transaccionals, usualment per entrada de dades i recuperació i processament de transaccions.

***Online.*** Fa referència a un estat de connectivitat, davant del terme fóra de línia (offline) que indica un estat de desconnexió.

***Open Source.*** És l'expressió amb la que es coneix al programari o maquinari distribuït i desenvolupat lliurement. Es focalitza més en els beneficis pràctics (accés al codi font) que en qüestions ètiques o de llibertat que tant es destaquen en el programari lliure.

***Plugin.*** És una aplicació que es relaciona amb una altra per agregar-li una nova funció i, generalment, molt específica. Aquesta aplicació addicional és executada per l'aplicació principal i interactuen per mitjà de la interfície de programació d'aplicacions.

***Reporting.*** És l'esforç de les empreses ens instrumentar processos, procediments i fluxos d'informació. El mètode d'obtenció d'informació derivat d'aquests fluxos d'informació han de ser el més ràpid, àgil i fiable possible.

***Script.*** És un arxiu d'ordres o de processament per lots que és interpretat per un intèrpret de comandes i s'utilitza per realitzar diverses tasques de forma seqüencial.

***SGBD.*** Sistema Gestor de Bases de Dades. És un conjunt de programes que permeten l'emmagatzematge, modificació i extracció de la informació a una base de dades, a més de proporcionar eines per afegir, esborrar, modificar i analitzar les dades. Els usuaris poden accedir a la informació utilitzant eines específiques d'interrogació i de generació d'informes, o bé mitjançant aplicacions a l'efecte.

***SQL (Strucutured Query Language).*** És un llenguatge declaratiu d'accés a bases de dades relacionals que permet especificar diversos tipus d'operacions en elles.

***Staging Area.*** És una àrea intermèdia d'emmagatzematge de dades utilitzada pel processament dels mateixos durant els processos d'extracció, transformació i càrrega (ETL). Aquesta àrea es troba entre la font de les dades i el seu destí, que moltes vegades són magatzems de dades, data marts o altres repositoris de dades.

# 7. Bibliografia

**Referencies web:**

https://www.gestiopolis.com/efectividad-en-la-publicidad-de-impacto/

https://larueding.com/2012/06/15/6-claves-para-el-uso-eficaz-de-los-visuales-en-la-publicidad/

http://www.redalyc.org/articulo.oa?id=99315569003

http://openaccess.uoc.edu/webapps/o2/handle/10609/39982 http://www.dataprix.com/que-es-un-datawarehouse

https://prezi.com/ft5pay1wbnjy/diferencias-entre-talend-pentaho-y-jasper/

https://www.yurbi.com/blog/best-open-source-business-intelligence-tools-for-tech- savvy-companies/

http://mondrian.pentaho.com/documentation/workbench.php

http://www.meteorite.bi/products/saiku

https://ruclip.com/video/JKbtf6tT5w4/procesos-etl-con-pentaho-sqlserver-postgresql-mysql-excel.html

http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ydirin_p_mm/capitulo2.pdf

http://www.dataintegration.info/etl

https://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm

http://www.sinnexus.com/business_intelligence/datawarehouse.aspx

http://bi-businessintelligence.blogspot.com.es/2009/05/jaspersoft.html

https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/qu-son-los-procesos-etl

http://openaccess.uoc.edu/webapps/o2/handle/10609/45341

http://openaccess.uoc.edu/webapps/o2/handle/10609/39982

# 8. Annexos

## 8.1 Annex 1. Script creació BD Staging Area

A continuació es mostren els scripts *staging_area.sql* emprats per la creació de la base de dades Staging Area.

**Staging_area.sql**

```
DROP TABLE IF EXISTS staging_area.visits;
DROP TABLE IF EXISTS staging_area.impacts;
DROP TABLE IF EXISTS staging_area.sales;
DROP TABLE IF EXISTS staging_area.points;
DROP TABLE IF EXISTS staging_area.products;

CREATE TABLE staging_area.points (
  id INT NOT NULL AUTO_INCREMENT PRIMARY KEY,
  zone varchar(20) NOT NULL,
  city varchar(20) NOT NULL,
  shopping_center varchar(40) NOT NULL,
  type varchar(20) NOT NULL
);

CREATE TABLE staging_area.products (
  id INT NOT NULL AUTO_INCREMENT PRIMARY KEY,
  family varchar(255) NOT NULL,
  product varchar(255) NOT NULL
);

CREATE TABLE staging_area.sales (
  id INT NOT NULL AUTO_INCREMENT PRIMARY KEY,
  id_point int NOT NULL,
  id_product int NOT NULL,
  sale_date datetime NOT NULL,
  sales double NOT NULL,
  FOREIGN KEY (id_point) REFERENCES staging_area.points(id),
  FOREIGN KEY (id_product) REFERENCES staging_area.products(id)
);

CREATE TABLE staging_area.impacts (
  id INT NOT NULL AUTO_INCREMENT PRIMARY KEY,
  id_point int NOT NULL,
  id_product int NOT NULL,
  impact_point varchar(30) NOT NULL,
  impact_date datetime NOT NULL,
  impacts double NOT NULL,
  FOREIGN KEY (id_point) REFERENCES staging_area.points(id),
  FOREIGN KEY (id_product) REFERENCES staging_area.products(id)
);
```

```
CREATE TABLE staging_area.visits (
  id INT NOT NULL AUTO_INCREMENT PRIMARY KEY,
  id_point int NOT NULL,
  visits_date datetime NOT NULL,
  visits double NOT NULL,
  FOREIGN KEY (id_point) REFERENCES staging_area.points(id)
);
```

## 8.2 Annex 2. Arxiu de definició dels cubs OLAP

L'arxiu *CubPubliDW.xml* és on es defineixen els dos cubs OLAP implementats des de l'aplicació *Schema Workbench.* Aquest fitxer conter la definició de les dimensions, jerarquies, nivells i mesures dels dos cubs.

```
<Schema name="PubliDW">
  <Cube name="Ventes" visible="true" cache="true" enabled="true">
    <Table name="fet_sales_impacts">
    </Table>
    <Dimension type="StandardDimension" visible="true" foreignKey="id_point" highCardinality="false" name="DimensioPunts">
      <Hierarchy name="JerarquiaPunts" visible="true" hasAll="true" primaryKey="id">
        <Table name="dm_points">
        </Table>
        <Level name="Tipus" visible="true" column="type" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Zona" visible="true" column="zone" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Ciutat" visible="true" column="city" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="CentreComercial" visible="true" column="shopping_center" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
```

```
<Dimension type="StandardDimension" visible="true" foreignKey="id_product" highCardinality="false" name="DimensioProductes">
  <Hierarchy name="JerarquiaProductes" visible="true" hasAll="true" primaryKey="id">
    <Table name="dm_products">
    </Table>
    <Level name="Familia" visible="true" column="family" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
    </Level>
    <Level name="Productes" visible="true" column="product" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
    </Level>
  </Hierarchy>
</Dimension>
<Dimension type="TimeDimension" visible="true" foreignKey="id_temps" highCardinality="false" name="DimensioTemps">
  <Hierarchy name="JerarquiaTemps" visible="true" hasAll="true">
    <Table name="dm_temps">
    </Table>
    <Level name="Any" visible="true" column="any" type="Integer" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="Estacio" visible="true" column="estacio" type="String" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="Trimestre" visible="true" column="trimestre" type="Integer" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="Mes" visible="true" column="mes" type="Integer" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="NomMes" visible="true" column="nom_mes" type="String" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="Dia" visible="true" column="dia" type="Integer" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
    <Level name="NomDia" visible="true" column="nom_dia" type="String" uniqueMembers="false" levelType="TimeYears" hideMemberIf="Never">
    </Level>
```

```
      <Level name="Data" visible="true" column="data" type="Timestamp" uniqueMembers="false"
levelType="TimeYears" hideMemberIf="Never">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaAny" visible="true" hasAll="true">
      <Table name="dm_temps">
      </Table>
      <Level name="Any" visible="true" column="any" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaTrimestre" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Trimestre" visible="true" column="trimestre" type="Integer"
uniqueMembers="false">
      </Level>
    </Hierarchy>
   <Hierarchy name="JerarquiaMes" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Mes" visible="true" column="mes" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaDia" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Dia" visible="true" column="dia" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
  </Dimension>
  <Measure name="Ventes" column="sales" datatype="Numeric" aggregator="sum" visible="true">
  </Measure>
  <Measure name="ImpactesBigScreen" column="impact_bigscreen" datatype="Numeric"
aggregator="sum" visible="true">
  </Measure>
  <Measure name="ImpactesPointScreen" column="impact_pointscreen" datatype="Numeric"
aggregator="sum" visible="true">
  </Measure>
 </Cube>
```

```
 <Cube name="Visites" visible="true" cache="true" enabled="true">
  <Table name="fet_visits">
  </Table>
  <Dimension type="StandardDimension" visible="true" foreignKey="id_point" highCardinality="false" name="DimensioPunts">
    <Hierarchy name="JerarquiaPunts" visible="true" hasAll="true">
      <Table name="dm_points">
      </Table>
      <Level name="Tipus" visible="true" column="type" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Zona" visible="true" column="zone" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
      <Level name="Ciutat" visible="true" column="city" type="String" uniqueMembers="false" levelType="Regular" hideMemberIf="Never">
      </Level>
    </Hierarchy>
  </Dimension>
<Dimension type="TimeDimension" visible="true" foreignKey="id_temps" name="DimensioTemps">
    <Hierarchy name="JerarquiaTemps" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Any" visible="true" column="any" type="Integer" uniqueMembers="false">
      </Level>
      <Level name="Estacio" visible="true" column="estacio" type="String" uniqueMembers="false">
      </Level>
      <Level name="Trimestre" visible="true" column="trimestre" type="Numeric" uniqueMembers="false">
      </Level>
      <Level name="Mes" visible="true" column="mes" type="Integer" uniqueMembers="false">
      </Level>
      <Level name="NomMes" visible="true" column="nom_mes" type="String" uniqueMembers="false">
      </Level>
      <Level name="Dia" visible="true" column="dia" type="Integer" uniqueMembers="false">
      </Level>
      <Level name="NomDia" visible="true" column="nom_dia" type="String" uniqueMembers="false">
      </Level>
      <Level name="Data" visible="true" column="data" type="Timestamp" uniqueMembers="false">
      </Level>
    </Hierarchy>
```

```
<Hierarchy name="JerarquiaAny" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Any" visible="true" column="any" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaTrimestre" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Trimestre" visible="true" column="trimestre" type="Integer"
uniqueMembers="false">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaMes" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Mes" visible="true" column="mes" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
    <Hierarchy name="JerarquiaDia" visible="true" hasAll="true">
      <Table name="dm_temps" alias="">
      </Table>
      <Level name="Temps" visible="true" column="dia" type="Integer" uniqueMembers="false">
      </Level>
    </Hierarchy>
  </Dimension>
  <Measure name="Visites" column="visits" datatype="Numeric" aggregator="sum" visible="true">
  </Measure>
 </Cube>
</Schema>
```