Universitat Oberta de Catalunya

uoc.edu

Sistema d'intel·ligència de negoci per a l'anàlisi de la publicitat en entorns digitals

Albert Ribó Pascual

Màster en Enginyeria Informàtica

Àrea de treball final: Business Intelligence

Consultor: David Amorós Alcaraz

Professora responsable de l'assignatura: Maria Isabel Guitart Hormigo

Data Lliurament: 11/06/2018

FITXA DEL TREBALL FINAL

Títol del treball:	*Sistema d'intel·ligència de negoci per a l'anàlisi de la publicitat en entorns digitals.*
Nom de l'autor:	*Albert Ribó Pascual*
Nom del consultor:	*David Amorós Alcaraz*
Nom del PRA:	*Maria Isabel Guitart Hormigo*
Data de lliurament (mm/aaaa):	*06/2018*
Titulació o programa:	*Màster en Enginyeria Informàtica*
Àrea del Treball Final:	*Business Intelligence*
Idioma del treball:	*Català*
Paraules clau	*Business Intelligence, Processos ETL, Anàlisi OLAP.*

Resum del Treball:

Vivim en una era en la que tot es digitalitza. El paradigmes coneguts fins ara ja no són útils i hem d'explorar noves alternatives. El món de la publicitat no n'és una excepció i les grans Xarxes Socials han incorporat al seu model de negoci plataformes de publicitat online. Aquestes plataformes, gràcies a l'entorn en el que es desenvolupen, poden proporcionar informació molt valuosa de forma quasi instantània. Aquest feedback d'informació ha de servir als anunciants per millorar el rendiment de les seves inversions en publicitat.

En aquest treball es mostrarà com, gràcies a eines de Business Intelligence, es pot ser capaç d'extreure relacions entre les dades per tal de millorar el rendiment dels anuncis. Per fer-ho s'estudiaran les plataformes BI, es triarà una i es seguiran els passos necessaris per obtenir la informació necessària i poder ajudar a la presa de decisions. Aquests passos són:

- Definició i implementació del Data Warehouse de la plataforma BI.
- Definició i implementació dels processos ETL.
- Implementació dels sistemes OLAP per explotar les dades del Data Warehouse.
- Anàlisi de les dades i obtenció de conclusions que ajudin a la presa de decisions.

Gràcies a l'explotació de les dades contingudes al DW, es podrà donar resposta a les preguntes analítiques formulades com a objectiu final d'aquest treball.

S'obtindrà una visió global de perquè les plataformes BI han esdevingut elements imprescindibles per a la competitivitat de les empreses, donant suport a l'anàlisi complex de dades multidimensionals per, finalment, guiar la presa de decisions basades en fets.

Abstract:

We live in an era in which everything is digitalized. The old paradigms are no longer useful and we have to explore new alternatives. The world of advertising is no exception and the main Social Networks have incorporated to their business model online advertising platforms. These platforms, thanks to the environment in which they are deployed, can provide very valuable information instantaneously. This feedback should be used by the advertisers to improve the performance of their advertising investments.

This paper will show how, thanks to Business Intelligence tools, we can be able to extract relationships between data in order to improve the performance of the ads. To do so, BI platforms will be studied, one will be chosen and all the necessary steps will be followed to obtain the necessary information and to be able to help with decision-making processes. These steps are:

- Definition and implementation of the Data Warehouse for the BI platform.
- Definition and implementation of ETL processes.
- Implementation of OLAP systems to exploit Data Warehouse data.
- Analysis of the data and obtaining conclusions for decision-making.

Thanks to the exploitation of the data contained in the DW, the analytical questions formulated as the final objective of this work can be answered.

A global vision will be obtained why BI platforms have become indispensable elements for the competitiveness of companies, supporting the complex analysis of multidimensional data to finally guide decision-making processes based on facts.

Índex

Llista de figures

1.Introducció

1.1 Context i justificació del Treball

Actualment la publicitat digital s'ha convertit en una de les principals fonts d'ingressos per a aquelles plataformes que n'exploten aquest model de negoci. Les xarxes socials més importants d'Internet també s'han sumat a aquest tipus de negoci, Facebook amb Facebook Ads, Twitter amb Twitter Ads, Youtube amb els vídeos TrueView (gestionats des de Google AdWords) i Instagram que gestiona la seva plataforma publicitària des de Facebook Ads; són un clar exemple.

Google i Facebook s'han convertit en els actors principals del mercat de la publicitat digital pràcticament arreu del món. Per a Facebook els ingressos reportats per la seva plataforma de publicitat digital es van situar al 2017 en 39,5 bilions de dòlars [1]. Google va obtenir uns ingressos d'aproximadament 95,4 bilions de dòlars durant l'any 2017 [2]. Molt lluny trobem a Twitter, que en 2017 va generar 2,1 bilions de dòlars en ingressos [3]. Els números de les principals plataformes de publicitat digital són increïbles, i les previsions per a l'any 2018 les situa en xifres d'ingressos rècord.

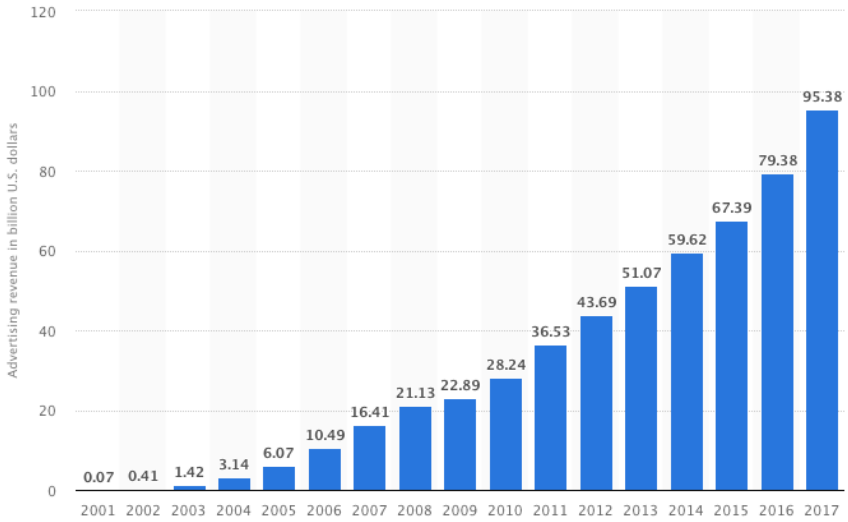


Figura 1: Ingressos publicitaris de Google del 2001 al 2017 (en bilions d'USD).

Aquestes xifres demostren com d'important és el màrqueting digital i les plataformes de publicitat que suporten aquest model de negoci. Des del punt de vista dels usuaris d'aquestes plataformes de publicitat, el model presenta clars avantatges envers les plataformes tradicionals. Ofereixen la possibilitat d'arribar a un públic focalitzat (global, nacional, local, etc.) permet segmentar les audiències en funció de sexe, gustos i preferències, nivell d'ingressos, lloc de residència (o punt d'accés a Internet), etc. Aquest gran ventall d'opcions mostra clarament el grandíssim potencial d'aquestes plataformes. Presenten un aparador de luxe per a totes aquelles empreses que es volen anunciar, ja sigui per captar a nous de clients, per consolidar la imatge de marca o per aconseguir objectius de conversió (vendes, subscripció a butlletins, etc.).

Tot i el gran potencial de les plataformes digitals, si les campanyes de publicitat no es dissenyen correctament, i sobretot si no es mesura constantment el seu rendiment, poden suposar una despesa molt gran per als anunciants. Donada la naturalesa de l'entorn, el feedback que proporcionen aquestes plataformes és pràcticament immediat (o quasi) de manera que l'avaluació del rendiment d'un anunci o campanya és pot fer ràpidament.

Aquesta immediatesa contrasta amb la lentitud de les plataformes clàssiques (TV, ràdio, premsa) on el mesurament és molt més complicat.

Qualssevol campanya publicitària que s'inicia persegueix uns objectius. Aquests objectius han de ser mesurables, per poder validar que les campanyes estan funcionant correctament i que hi ha un retorn de la inversió (ROI). Aquest ROI es calcularà en base als objectius de la campanya i unes vegades informarà sobre els ingressos obtinguts en vendes de productes, d'altres informarà dels clients registrats en un web site, d'altres mesurarà el transit generat cap a una botiga online, etc.

La capacitat de focalitzar les campanyes (o segmentar-les) segons l'audiència objectiu, ens ofereix la possibilitat d'analitzar com canvis en els paràmetres de les campanyes afecten al rendiment final dels anuncis. Les plataformes de publicitat digital generen una quantitat molt gran de dades que poden ser utilitzades i analitzades per sistemes d'intel·ligència empresarial (Business Intelligence) per tal de poder prendre decisions beneficioses per a la companyia. El medi on es publica l'anunci (Internet) facilita l'obtenció d'una gran quantitat de dades relacionades amb l'anunci, i amb informació personal del visitant que ha clicat a l'anunci. Aquesta gran quantitat d'informació heterogènia (de diverses fonts i en diversos formats) ha de ser emmagatzemada de tal forma que permeti un accés senzill i ràpid a la informació continguda. Els sistemes de BI ens proporcionen una plataforma consolidada per realitzar aquestes accions.

En aquest TFM es vol donar suport a la decisió sobre quin o quins paràmetres, o quina combinació d'aquests paràmetres, generen més benefici per a l'Empresa (millor efectivitat) a partir de les dades recollides en diferents xarxes socials. El principal indicador (KPI - Key Performance Indicator) que s'utilitzarà per mesurar l'efectivitat dels anuncis serà el CTR (Click Through Rate). Aquest indicador ens informa del percentatge de clics en un anunci per impressions del mateix, i es transforma en un percentatge, que es pot interpretar segons la campanya publicitària. Per exemple, si l'objectiu és atraure transit a un lloc web, interessarà obtenir un CTR alt, que ens indicarà que l'anunci funciona de forma correcta ja que el públic objectiu clica l'anunci cada poques impressions de l'anunci per pantalla. En aquest treball prendrem el CTR com l'indicador que mesura l'efectivitat d'un anunci.

1.2 Objectius del Treball

L'objectiu final del TFM és donar resposta a una sèrie de qüestions rellevants a partir de l'explotació de les dades generades en les diferents plataformes de publicitat de les principals xarxes socials. Aquestes respostes s'obtindran gràcies a l'ús d'un sistema de BI. Aquest sistema BI s'haurà de definir, modelar i implementar.

Les respostes obtingudes a les qüestions formulades, permetran confirmar si hi ha una relació entre la segmentació del anuncis i el seu rendiment i, a més, si existeix una relació amb el tipus d'anunci publicat. La formulació de les qüestions es mostra a continuació:

- Quines regions o ciutats tenen millors indicadors d'efectivitat? Hi ha alguna relació amb el producte o família de productes?
- Existeixen una relació entre la millora dels indicadors d'efectivitat amb algun segment de la població objectiu?
- Hi ha alguna plataforma on, sota les mateixes condicions, s'obtinguin millors taxes de visualització?
- Existeixen relacions entre plataformes i franges d'edat d'usuaris que provoquin millors taxes de visualització?
- El coneixement del grup d'interès dels usuaris podria ajudar a millorar els indicadors per determinats productes?

Per poder respondre de forma correcta a aquestes preguntes necessitem un sistema que ens doni suport per a l'anàlisi de la informació. Per aquest motiu, com s'ha comentat

anteriorment, s'hauran d'assolir uns objectius previs relacionats amb l'anàlisi, creació, injecció de dades i explotació de les mateixes en un entorn de Business Intelligence. Abans d'arribar a les conclusions s'haurà d'haver assolit les fites següents:

- Anàlisi de sistemes de BI que puguin donar suport a la resolució del nostre problema. Aquest anàlisi ens conduirà a la tria de la plataforma de BI que s'emprarà durant la resta del TFM.
- Disseny i implementació del Data Warehouse que proporcioni el suport per tal de poder explotar les dades recopilades a les plataformes publicitàries de les principals xarxes socials.
- Implementació dels processos ETL (Extracció, Transformació i Càrrega) de les dades des de formats heterogenis cap al Data Warehouse per tal de deixar la informació rellevant estructurada per poder ser explotada i analitzada.
- Explotació de la informació del Data Warehouse emprant-les eines que es considerin oportunes (per exemple cubs OLAP), i presentació de la informació en un format, clar, comprensible i útil per als usuaris que han d'analitzar-la i prendre decisions en funció de la informació presentada.

1.3 Enfocament i mètode seguit

El mètode seguit en la realització d'aquest TFM, es correspon amb les fases essencials en el desenvolupament de projectes de BI. Podem identificar aquestes fases com:

- Anàlisi del problema a resoldre
- Selecció de la plataforma BI que cobreixi les necessitats
- Disseny i implementació del Data Warehouse
- Disseny i implementació dels processos ETL
- Explotació de dades
- Obtenció i interpretació dels resultats.

Per al desenvolupament del TFM, caldrà utilitzar una plataforma de BI que ens proporcioni totes les eines necessàries per assolir els objectius. Al mercat hi ha una gran quantitat de plataformes BI (tant propietàries com Open Source), per tant es triarà d'entre totes les opcions aquella que encaixi millor segons les necessitats del problema a ressoldre.

Sobre la plataforma BI es realitzaran els desenvolupaments necessaris (processos ETL, explotació de dades, etc.) per cobrir els requeriments del problema a resoldre i finalment s'obtindran i analitzaran els resultats

1.4 Planificació del Treball

La planificació del TFM que que s'ha realitzat, presenta 7 fases que alhora es divideixen en tasques. A continuació es mostra la llista de fases i tasques, completada amb una descripció dels treballs a realitzar en cadascuna de les fases (i tasques principals):

1. *Realització de la memòria final*. La primera fase és la elaboració de la memòria final del TFM. Aquesta fase es du a terme durant tot el temps d'execució del TFM, i es realitzarà en paral·lel als diferents treballs.
2. *Planificació*. És la fase inicial del TFM, en la que s'estableixen els objectius a assolir, la planificació i la divisió en tasques. Aquesta fase finalitza el 05/03/2018 i culmina amb la entrega del document de la PAC1.

3. *Anàlisi d'entorns de Business Intelligence*. Durant aquesta fase es realitzaran tasques enfocades a l'anàlisi de plataformes de BI disponibles al mercat. S'avaluaran aspectes com: tipus de llicència necessària, requeriments de HW de la plataforma, funcionalitats i prestacions que ofereixen, etc. En funció de tota la informació recopilada es triarà la plataforma sobre la que desenvolupar les fases següents de TFM. És una fase curta que té una durada aproximada d'una setmana.
   Aquesta fase finalitza aproximadament el 13/03/2018.
4. *Disseny del Data Warehouse.* És la primera de les tres fases de llarga durada del TFM. En aquesta fase es realitzaran diverses tasques que tindran com a objectiu modelar el Data Warehouse. Les tasques a realitzar són les següents:
   - Anàlisi de les dades del problema a resoldre. Analitzarem les dades d'entrada del nostre problema per conèixer de quina informació disposem i com està organitzada. Aquesta informació en format heterogeni acabarà dins de Data Warehouse.
   - Anàlisi dels possibles dissenys del data Warehouse. Recopilació d'informació dels models de disseny de Data Warehouse, avaluació i finalment, tria del disseny més convenient per resoldre el nostre problema
   - Implementació del model. Un cop s'ha triat el model de disseny (p. ex.: estrella o snowflake), es realitzarà la modelització o implementació efectiva en la plataforma de BI. Aquesta implementació estarà fonamentada pels objectius a assolir.

   Com a resultat d'aquesta fase obtindrem la implementació del Data Warehouse. Arribat a aquest punt el nostre Data Warehouse estarà a punt per ser utilitzat, tot i que encara no tindrà dades. Aquesta fase té una duració aproximada de tres setmanes, finalitzant el 03/04/108; i culmina amb l'entrega de la PAC2, el 09/04/2018, on s'inclourà tota la informació relacionada amb les fases 3 i 4. Entre d'altres els elements lliurables a la PAC2 seran el document comparatiu de les diferents plataformes de BI, justificació de la plataforma seleccionada, documents de disseny del Data Warehouse, etc.

5. *Disseny dels processos ETL.* Segona fase de llarga durada (4 setmanes aproximadament). Les tasques d'aquesta fase es centren en la definició dels processos ETL per tal de traspassar les dades des de les fonts heterogènies al Data Warehouse. El detall de les tasques a realitzar és el següent:
   - Definició i implementació dels processos ETL a implementar. La implementació del processos ETL és dependent de la plataforma de BI que es triï en la fase 3. La definició i implementació, per tant, serà depenent de plataforma i es tindrà que fer en funció de les eines i els mètodes que aquesta ens proporcioni.
   - Injecció de dades al Data Warehouse des dels processos ETL. Un cop els processos ETL estan definits i implementats, s'inicia la fase en la que es realitza la injecció d'informació des de les fonts originals fins al Data Warehouse. Durant aquesta fase s'hauran de realitzar totes les validacions necessàries per assegurar que les dades s'han injectat correctament al Data Warehouse.

   Al finalitzar aquesta etapa el Data Warehouse estarà preparat per iniciar els processos d'explotació de dades. La finalització d'aquesta fase està planificada per al 02/05/2018. La entrega de la PAC3, el 07/05/2018, recollirà tota la informació rellevant d'aquesta fase. Els elements lliurables d'aquesta fase seran (entre d'altres) els documents explicatius dels processos ETL, els documents tècnics relacionats amb la implantació dels processos ETL, etc.

6. *Explotació de les dades*. Aquesta fase es planifica amb una durada aproximada de tres setmanes finalitzant el 27/05/2018. Durant aquesta fase es realitzaran les tasques necessàries per explotar la informació continguda al Data Warehouse (per

exemple amb cubs OLAP). La divisió en tasques d'aquesta fase és mostra a continuació:
    - Anàlisi i tria del(s) mètode(s) d'explotació de dades. S'analitzaran les opcions d'explotació de dades existents; i es triarà la que millor s'adapti per resoldre satisfactòriament el problema plantejat.
    - Representació visual de dades en funció del mètode d'explotació triat. Generació d'informes, quadres de comandament, gràfiques o llistats necessaris per tal de poder respondre a les preguntes analítiques formulades.
    - Obtenció de respostes a les qüestions analítiques formulades. Amb la informació obtinguda en la tasca anterior estarem en disposició de donar respostes a les preguntes analítiques.
7. *Anàlisi dels resultats i conclusions*. La darrera fase es centra en la elaboració d'un resum detallat dels resultats i de les conclusions que s'obtenen de la informació mostrada pel sistema de BI.
   Aquesta fase finalitza el 10/06/2018, i culmina amb la entrega de la memòria final del TFM que contindrà tota la informació del TFM així com els elements lliurables que es determinin com a necessaris; com per exemple documents de modelització del Data Warehouse, la implementació dels processos ETL, etc.

Un cop arribat aquest punt, es podrà lliurar la memòria final del TFM, juntament amb tots els artefactes lliurables per a la seva avaluació.

La planificació en format de diagrama de Gantt es pot consultar a l'annex 1 del present document.

1.5 Breu sumari de productes obtinguts

La elaboració d'aquest TFM generarà els següents productes principals:

1. Disseny del Data Warehouse
2. Disseny dels processos ETL de càrrega del Data Warehouse
3. Processos d'explotació de dades (cubs OLAP)
4. Anàlisi de resultats i conclusions a partir de la informació proporcionada per la plataforma de BI.

1.6 Breu descripció dels altres capítols de la memòria

A continuació es presenta una breu descripció del contingut de la resta de capítols que componen aquesta memòria:

- *Capítol 2: Anàlisi d'entorns de Business Intelligence*. En aquest capítol es presenta una breu descripció de les plataformes de BI, quin és el seu origen i quines opcions hi ha actualment al mercat. Finalment es comparen les principals plataformes de BI per finalment triar la plataforma sobre la que es desenvoluparan els treballs d'aquest TFM.
- *Capítol 3*: *Disseny del Data Warehouse*. En aquest capítol s'exploren les diferents opcions de disseny del DW, analitzant els seus avantatges i inconvenients, per finalment obtenir un disseny del DW que compleixi amb els requeriments del problema a solucionar.
- *Capítol 4*: *Modelització i implementació dels processos ETL*. Donat el disseny del Data Warehouse presentat en el capítol anterior, es modelitzen els processos ETL

que serviran per transformar les dades d'origen del cas d'estudi i carregar-les en el format correcte al magatzem de dades. Addicionalment es presentarà la implementació d'aquests processos en l'eina ETL de la plataforma BI triada per a la elaboració d'aquest TFM.

- *Capítol 5*: *Disseny i implementació de cubs OLAP per a l'explotació de la informació del DW.* Aquest capítol presenta una descripció dels diferents tipus de cubs OLAP que es poden implementar per a l'explotació de les dades d'un DW, així com la implementació efectiva del cub OLAP que servirà per poder analitzar les dades del DW del probelama a estudiar i donar resposta a les preguntes analítiques.
- *Capítol 6: Anàlisi de dades del DW a partir de cubs OLAP*. Aquest capítol conté l'anàlisi de dades realitzar a partir de cubs OLAP i l'eina Saiku Analitycs per tal de donar resposta a les qüestions analítiques formulades.
- *Capítol 7: Conclusions*. Per finalitzar aquest TFM, es recullen totes les conclusions obtingudes durant la realització del projecte.

2. Anàlisi d'entorns de Business Intelligence

2.1 Breu historia del Business Intelligence

L'aparició del concepte Business Intelligence data de l'any 1865, on Richard Millar Devens el va fer servir a "C*yclopædia of Commercial and Business Anecdotes"*, per referir-se com el banquer Sir Henry Furnese, s'aprofitava de la informació que havia pogut anar recollint al llarg del temps per tal d'aconseguir un avantatge estratègic enfront els seus competidors comercials. Al 1958 l'investigador d'IBM Hans Peter Luhn, descriu en l'article "A Business Intelligence System" el BI com "l'habilitat d'aprendre les relacions de fets que ja han ocorregut per tal de guiar properes accions cap a una meta desitjada". En aquest article queda pales el potencial de recollir grans quantitats d'informació per obtenir avantatges empresarials a través de l'ús de la tecnologia.

El concepte modern de BI, data de fa 60 anys, i actualment l'entenem com l'ús de la tecnologia per tal de recollir i recopilar grans quantitats d'informació, analitzar-la i transformar-la en informació útil per poder actuar en un breu període de temps. Aquesta forma d'entendre el BI va lligada de forma necessària a la tecnologia, ja que només gràcies a ella som capaços de realitzar els processos esmentats anteriorment. Recordem que no es fins al 1970 que no apareixen les bases de dades relacionals, que van facilitar l'emmagatzemament i accés a grans quantitats de dades i que van ser adoptades de forma global.

Però a mesura que la tecnologia avançava i les necessitats empresarials eren més exigents, les Bases de Dades Relacionals ja no complien tant bé amb les necessitats dels negocis. Va ser en la dècada dels 80, on autors com Bill Imon i Ralph Kimball creen el concepte de Data Warehouse, entès com un gran magatzem de dades centralitzat Durant aquells anys l'alt cost de la tecnologia feia que només grans corporacions amb molts recursos poguessin disposar de sistemes de BI.

A la dècada dels 90, tot i que les solucions de BI seguien sent molt costoses, van néixer una gran quantitat d'empreses especialitzades en aquesta àrea de la computació i el concepte de BI va començar a ser molt conegut arreu del món, és l'època en que el BI es dona a conèixer a nivell mundial. Cal tenir en compte que degut a les limitacions de la tecnologia obtenir "respostes" dels sistemes de BI d'aquella època podia trigar dies (fins i tot setmanes) i per tant les corporacions que disposaven de sistemes de BI escollien especialment aquelles preguntes de les que millor rendiment en podien treure. A més aprofitaven hores de poca càrrega als sistemes (generalment de matinada) per llançar els processos de BI.

A partir de l'any 2000 Internet esdevé popular i el PC es troba ben assentat entre els usuaris particulars. Tecnològicament, el cost de producció de components es redueix en la mesura que la potència de càlcul i capacitat d'emmagatzemament augmenta. És el moment en el que les plataformes BI esdevenen més senzilles d'usar (ja no cal personal especialitzat) i molt més flexibles. L'abaratiment de la tecnologia i l'aparició de plataformes BI Open Source, fa que les petites i mitjanes empreses puguin accedir a aquestes plataformes, ja sigui amb versions Open Source, o a través de sistemes propietaris amb un cost ajustat i prestacions suficients per assolir els seus objectius. El BI esdevé popular a nivell empresarial oferint solucions adaptades tant a grans corporacions com a petites i mitjanes empreses.

Actualment, el ritme vertiginós en el que la tecnologia ha avançat fa que cada cop es generin més i més dades des de fonts totalment heterogènies, com per exemple la quantitat de dades (rellevants o no) que generen les xarxes socials (Facebook, Instagram, LinkedIn, etc.). La necessitat d'emmagatzemar, processar i extreure la informació rellevant ha generat nous camps tecnològics lligats amb el BI com: Big Data o Business Analytics.

El BI és una disciplina de la computació fortament arrelada al món empresarial i actualment tota mena d'empreses, des de grans corporacions passant per mitjanes empreses i arribant a petits i modestos comerços online, fan servir sistemes de BI per tal de poder emprar totes

les dades que tenen a la seva disposició per guiar de forma correcta les decisions, obtenint avantatges competitius en un món que avança i genera dades a un ritme imparable.

2.2 Tipus de plataformes de Business Intelligence

En el moment d'escollir una plataforma BI, s'ha de realitzar una tasca d'anàlisi per avaluar els avantatges i inconvenients de cadascuna d'elles, verificar que les necessitats de l'organització queden satisfetes, que la implantació és possible (tècnicament i econòmicament), per finalment escollir de forma raonada la millor opció.

Històricament les plataformes de BI eren solucions de software empresarial molt cares i amb un cost d'implantació i manteniment molt elevat. Cal recordar que la majoria de proveïdors oferien paquets tancats i amb un cost de llicència propietària molt alt. Per aquest motiu només les grans corporacions es podien permetre sistemes de BI. Actualment el panorama ha canviat i podem trobar les plataformes propietàries, plataformes Open Source i plataformes i serveis al núvol.

L'aparició de plataformes Open Source ha permès que petites i mitjanes empreses amb pressupostos més ajustats, accedeixin a aquestes solucions, millorant la seva competitivitat. En aquest sentit també ha sigut un gran avenç la possibilitat de disposar plataformes a Internet, que s'aprofiten de sistemes distribuïts oferint el software com a servei (SaS) permetent així reduir els costos d'infraestructura i implantació de la plataforma a les empreses.

Els components que les plataformes completes de BI han de contenir es poden veure a la figura següent:

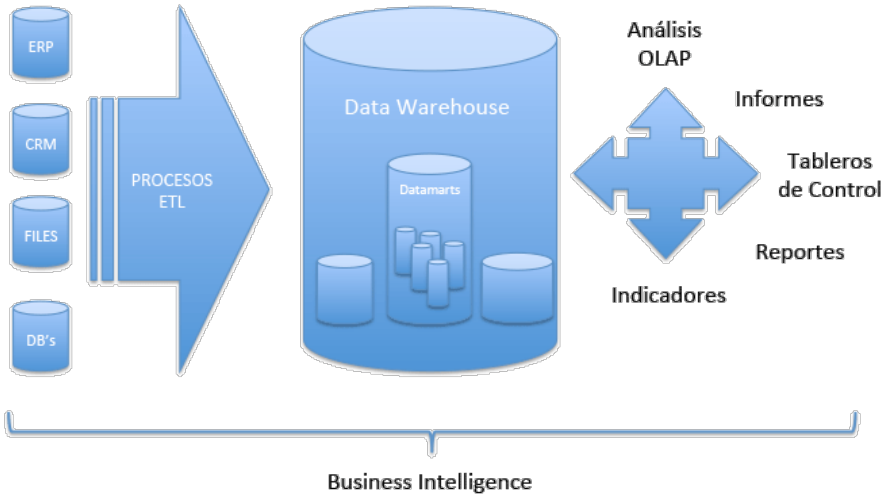


Figura 2: Components de les plataformes BI

Aquests component s'agrupen en quatre blocs:

1. Fonts d'informació: Són tots aquells elements del sistema empresarial que contenen dades de l'empresa, que poden ser utilitzades en un sistema BI. Ens trobem amb fitxers heterogenis de dades (per exemple fitxers Excel), Bases de Dades, altres sistemes corporatius, etc.
2. Processos ETL: Es corresponen a totes les activitats de transformació i processament de les dades de les fonts d'informació cap al magatzem de dades de la plataforma BI. En general les suites de BI proporcionen eines gràfiques que faciliten les tasques de disseny i implementació dels processos ETL.

3. Data Warehouse: És el magatzem de dades del sistema BI. És el lloc on estarà tota la informació empresarial susceptible de ser consultada per l'empresa.
4. Eines de visualització i explotació de la informació: En aquest bloc ens trobem un conjunt d'eines que permeten la consulta complexa i multidimensional de les dades (OLAP) juntament amb eines de visualització (dasboards, llistats, etc.), reporting i de descobriment (mineria de dades).

Al mercat hi ha diferents opcions i configuracions d'eines BI. Es important conèixer quines operacions es poden realitzar sobre una eina concreta per tal de poder realitzar l'anàlisi de forma correcta. En aquest sentit hem de diferenciar entre suites completes de BI i eines de BI que es centren en una part concreta. Com a exemple de suite BI completa podem referir-nos a SAP BuisnessObjects (oferint una gran quantitat d'opcions i funcionalitats) amb totes les eines necessàries de BI en el mateix entorn (sense necessitat de software de tercers) En l'altre extrem (eines que es centren en un aspecte concret del BI) trobem a *Tableau.* Aquesta eina permet la visualització d'informació de forma interactiva amb les dades d'un DW, però no es pot considerar com a plataforma completa de BI al no oferir una eina (dins del paquet) que permeti modelar i implementar processos ETL.

A continuació es presenten les plataformes BI propietàries i Open Source amb més acceptació al mercat.

2.2.1 Plataformes Business Intelligence propietàries

En el sector de les plataformes BI propietàries hi ha múltiples proveïdors que proporcionen paquets de software complerts amb una gran quantitat de funcionalitats i que estan dirigits a un espectre ampli de clients, adaptant la parametrització del software de BI a les peculiaritats dels clients. Alguns dels proveïdors de suites BI més importants del mercat són els següents:

- SAP amb SAP BusinessObjects (SAP BI/BW). SAP és una companyia d'origen alemany fundada al 1969 per enginyers d'IBM. En 2008 SAP adquireix a Business Objects, empresa especialitzada en intel·ligència de negoci. La plataforma ofereix versions tant on-premise com cloud [9]. La suite és altament parametritzable per adaptar-se a necessitats molt variades. El cost de llicències SAP BI/BW és realment alt, i la inversió en hardware i infraestructura també s'ha de tenir en compte.
- Oracle Corporation amb Oracle BI: Oracle és una companyia nord-americana fundada al 1977. La plataforma BI d'Oracle presenta totes les característiques de les plataformes BI completes. Al ser una plataforma propietària la integració es fa amb els productes propis d'Oracle, per exemple la base de dades [10].
- IBM amb IBM Cognos. IBM és una multinacional nord-americana fundada al 1911. En 2008 adquireix l'empresa Cognos especialitzada en intel·ligència de negoci. La plataforma ofereix funcionalitats cloud [11]. És una solució molt competitiva però que presenta dificultats d'implantació i manteniment.
- Qlik Technologies amb QlikView. Qlik és una empresa sueca fundada al 1993 i que està especialitzada en BI. Ofereix un paquet complert de BI, amb un preu de llicències superior al de la resta de competidors. Tot i oferir una plataforma completa una de les mancances més importants és que no suporta cubs OLAP. [12][13]
- Microsoft amb Microsoft Power BI. Microsoft és una multinacional nord-americana fundada el 1975. Power BI és una eina de BI senzilla d'usar, que ofereix connectivitat amb múltiples orígens de dades i un preu molt competitiu (amb versió gratuïta limitada). És una alternativa a tenir en compte però no ofereix totes les característiques de plataformes BI completes.[14][15]

Com es pot observar, dins dels proveïdors de solucions BI propietàries trobem algunes de les companyies de software més importants del món. Algunes tenen un llarg recorregut i d'altres han adquirit empreses especialitzades en solucions BI per entrar a competir en aquest segment del mercat del software.

Com es pot observar en la figura següent, els líders del mercat (del 2017) en el segment del BI es corresponen amb plataformes propietàries:

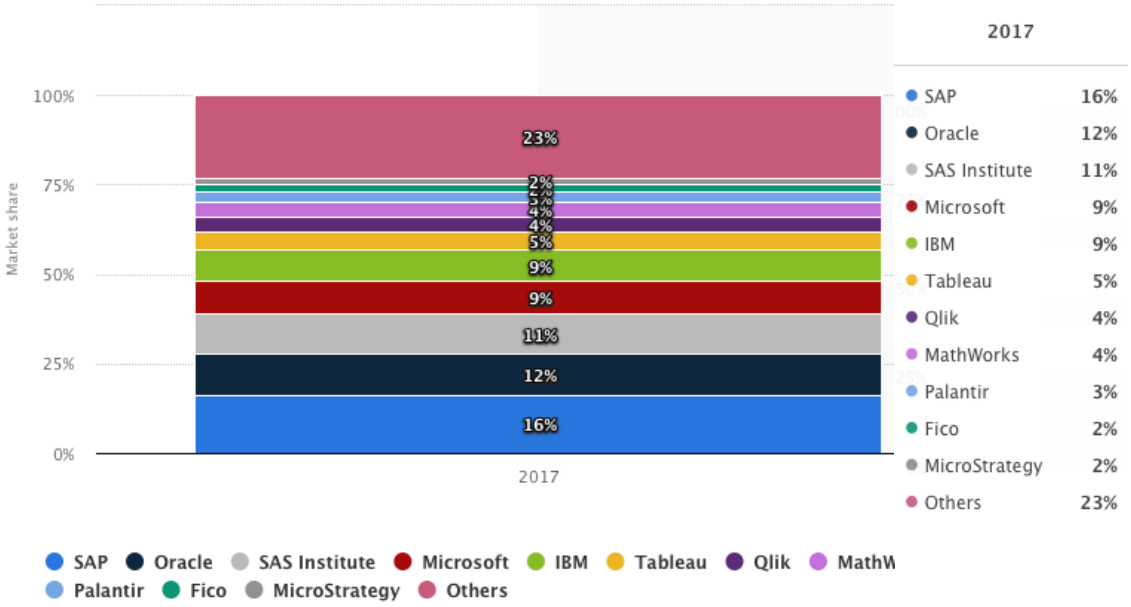


Figura 3: Quota de mercat de proveïdors de BI (2017)

Un dels gran inconvenients de les plataformes propietàries (deixant de banda l'elevat cost de les llicències en algunes d'elles) són els requeriments de HW, d'infraestructura (servidors dedicats, infraestructura de xarxa adequada, etc.) i manteniment especialitzat necessaris per al seu correcte funcionament.

2.2.2 Plataformes Business Intelligence Open-Source

El mercat de les plataformes de BI Open Source presenta menys opcions sobre les que triar. Tot i això hi ha suites BI interessants:

- Pentaho: Pentaho, originalment era una empresa nord-americana especialitzada en software de BI que en 2015 va ser adquirida per Hitachi Data Systems. Pentaho ofereix la seva plataforma en versió Open Source (Community Edition) i versió empresarial llicenciada (Enterprise Edition). La versió empresarial conté funcionalitats addicionals i suport oficial a la plataforma. Proporciona una plataforma BI completa (processos ETL, OLAP, etc.) [16][17][21]
- BIRT: El projecte "Business Intelligence and Report Tools" (BIRT) neix al 2004 dins del projecte Eclipse. Proporciona una solució completa de BI des dels reports operacionals fins als cubs OLAP. Tot i que amb components addicionals, podem considerar BIRT com una plataforma sencera, el seu enfocament principal està en la generació d'informes i visualització de dades per ser incrustades dins d'altres aplicacions.[16][18][22]
- Jaspersoft amb Jaspersoft Community: fundada a l'any 2001 i adquirida el 2015 per TIBCO, presenta una plataforma completa de BI Open Source. La plataforma està

composada per diversos components: Jaspersoft Server (servidor d'informes, gestió OLAP), Jaspersoft ETL (processos ETL), Jaspersoft Studio (disseny d'informes). TIBCO Jaspersoft també proporciona una versió llicenciada de la plataforma amb accés a suport tècnic, formacions, serveis de consultoria, etc.[16][23]

- Knowage: Es la darrera versió de la plataforma BI SpagoBI. Knowage facilita dos tipus de versions, la versió Community (Open Source) i la versió Enterprise (de pagament). La versió Open Source és plenament funcional i integra totes les eines d'una plataforma completa de BI. La versió empresarial conté funcionalitats addicionals que proporcionen valor afegit a la plataforma (assessorament, serveis de consultoria, etc.). [16][20][24]

Dins de les suites Open Source, s'ofereixen opcions de garanties per afrontar projectes BI complexos. Hi ha productes amb un ampli recorregut, i que ofereixen garanties de funcionament i de resultats. Posar a disposició d'empreses (generalment petites i mitjanes) eines Open Source permet que aquestes adquireixin coneixements més amplis del mercat i dels processos interns, podent-se transformar en empreses més competitives.

Plataformes completes de BI com Knowage (antic SpagoBI) i Pentaho, són candidats a ser tinguts en compte en el procés de selecció de plataforma BI, fins i tot quan hi ha candidats amb llicències propietàries. L'experiència en sector, la qualitat del software proporcionat i la característica de que són plataformes completes (sense necessitat d'eines externes a la plataforma per realitzar tasques; per exemple: processos ETL), els posiciona dins de l'elit de sistemes BI.

Una demostració de la fortalesa d'aquestes plataformes Open Source el trobem amb Pentaho i les seves mencions al "Gartner Magic Quadrant for Business Intelligence" [27].

2.3. Plataforma Business Intelligence triada per al desenvolupament del TFM

Per al desenvolupament del TFM, s'ha triat la plataforma de BI Pentaho en la seva versió 7 Community Edition (a data d'aquest treball la darrera versió disponible de la plataforma és la 8). Pentaho ofereix una plataforma completa de BI que cobreix tots els requeriments necessaris d'aquest TFM.

No s'ha triat una plataforma propietària perquè les versions d'avaluació limiten el seu ús a un període curt de temps o presenten limitacions de funcionalitats. Aquests condicionants ens orienten cap a una plataforma Open Source.

Pentaho és una suite completa de BI, que en funció de les necessitats permet la instal·lació de mòduls independents. Notar que, Pentaho, també ofereix una opció de pagament amb funcionalitats addicionals o ampliades, però que no són rellevants o necessàries per assolir els objectius d'aquest treball. La versió CE de Pentaho ofereix (ja sigui per defecte o amb complements Open Source addicionals) totes les funcionalitats necessàries: connectivitat a fonts d'informació, definició i implementació de processos ETL, disseny, implementació i accés al DW, eines OLAP per poder explotar la informació i eines per a la visualització de la informació (informes, quadres de comandament, etc.)

Pentaho és una solució d'ampli recorregut en el mercat, amb un bon suport de la comunitat (fòrums, manuals, etc.) i que sembla que té el seu futur com a plataforma BI garantit (versió Enterprise de pagament i versió Community gratuïta). Aquesta trajectòria com a solució BI, ofereix garanties que el software funcionarà de forma correcta sense tenir que realitzar accions addicionals (modificar fitxers font, fitxers de configuració, etc.). Com que la comunitat que hi ha al darrere és nombrosa, qualsevol dubte o problema que aparegui podrà ser consultat a Internet.

Els requeriments de HW necessaris per poder executar la plataforma no són exigents, i amb un ordinador personal es pot executar sense inconvenients. Cal remarcar que per

aplicacions amb un volum molt gran de dades i uns requeriments estrictes, caldria disposar de HW adequat per a la realització de les tasques. Un altre punt favorable és que Pentaho presenta versions per Windows, Linux i macOS.

Pentaho acumula una gran cartera de clients que confien en la plataforma [7] com: Telefonica, Nasdaq, Ranstad, British Telecom (BT), Zalando, etc.

S'ofereix connectivitat amb gran quantitat de motors de BD, tant propietaris com Open Source. En aquest sentit i per a la realització d'aquest treball s'ha triat MySQL. MySQL és un SGBD que ofereix versió Open Source (i llicències propietàries gestionades per Oracle), porta anys al mercat i ha demostrat la seva versatilitat i rapidesa en entorns exigents, com per exemple entorns Web. A més es disposen de múltiples eines de gestió gràfiques del SGBD que permetrà gestionar les dades còmodament.

MySQL és un SGBD que es utilitzat per algunes de les companyies tecnològiques més importants [8] a nivell global com: Facebook, Netflix, Google, Twitter, Youtube, Cisco, etc.

3. Disseny del Data Warehouse

3.1 Anàlisi de les dades del cas d'estudi

Un cop decidida la suite BI que s'utilitzarà durant el TFM, és el moment d'iniciar el procés de disseny de la peça angular de la solució BI, el Data Warehouse. El magatzem de dades requereix ser dissenyat amb cura i de forma correcta. A nivell corporatiu es converteix en una eina bàsica per la pressa de decisions dels processos de negoci i ha de ser accessible a gran velocitat (les consultes han de retornar la informació en menor temps possible).

El disseny de DW ha de ser pensat en funció de les dades que contindrà, i la informació que s'espera obtenir d'elles, tanmateix s'han de tenir en compte les operacions de *"Drill Down"* per aprofundir en les dades (obtenir un nivell de detall major) i *"Roll Up"* per obtenir una visió més global (obtenir una visió més general).

Les dades bàsiques del nostre cas d'estudi es proporcionen en un fitxer d'Excel. Aquestes dades es troben agrupades per pestanyes. A continuació es detalla el contingut del fitxer Excel:

- *Pestanya PRODUCTS*: Es troba la informació essencial dels productes que s'anuncien a les plataformes digitals. Hi ha un total de 8 productes diferents. Les dades de la pestanya són:
    - *Product*: Nom del producte.
    - *Family*: Família a la que pertany un producte.
- *Pestanya ZONES*: Es troba la informació essencial de les zones geogràfiques on s'han mostrat els anuncis. Hi ha un total de 16 zones diferents. Les dades de la pestanya són:
    - *Zone*: Nom de la zona.
    - *City*: Ciutat dins de la zona.
    - *ZipCode*: Codi postal de la ciutat.
- *Pestanya INSTAGRAM*: Es troba la informació agrupada de les visualitzacions d'anuncis en la xarxa social Instagram. Cada fila de la pestanya es correspon a un snapshot diari de les publicacions i clics en un anunci agrupat per: data, codi postal, producte, rang d'edat del visualitzador, sexe del visualitzador i aficions del visualitzador. Les dades de la pestanya són:
    - *Date*: Data de la observació.
    - *ZipCode*: Codi postal des del que s'ha visualitzat l'anunci.
    - Product: Producte que s'anuncia.
    - *Age*: Rang d'edat de l'usuari que ha visualitzat l'anunci. (4 rangs d'edat diferents).
    - *Gender*: Gènere (sexe) de l'usuari que ha visualitzat l'anunci. (2 gèneres diferents).
    - *Likes*: Aficions de l'usuari que ha visualitzat l'anunci. (8 aficions diferents).
    - *Prints*: Número de vegades que s'ha mostrat l'anunci en funció dels factors descrits anteriorment (data, codi postal, producte, etc.).
    - *Hits*: Número de vegades que s'ha clicat en l'anunci en funció dels factors descrits anteriorment (data, codi postal, producte, etc.).

- Pestanya FACEBOOK: Es troba la informació agrupada de les visualitzacions d'anuncis de la xarxa social Facebook. Cada fila de la pestanya es correspon a un snapshot diari de les publicacions i clics en un anunci agrupat per: data, codi postal, producte, rang d'edat del visualitzador, sexe del visualitzador i aficions del visualitzador. Les dades de la pestanya són les mateixes que per Instagram.
- Pestanya YOUTUBE: Es troba la informació agrupada de les visualitzacions d'anuncis de la xarxa social Youtube. Cada fila de la pestanya es correspon a un snapshot diari de les publicacions i clics en un anunci agrupat per: data, codi postal, producte, rang d'edat del visualitzador, sexe del visualitzador i aficions del visualitzador. Les dades de la pestanya són les mateixes que per Instagram.
- Pestanya TWITTER: Es troba la informació agrupada de les visualitzacions d'anuncis de la xarxa social Twitter. Cada fila de la pestanya es correspon a un snapshot diari de les publicacions i clics en un anunci agrupat per: data, codi postal, producte, rang d'edat del visualitzador, sexe del visualitzador i aficions del visualitzador. Les dades de la pestanya són les mateixes que per Instagram.

Arribats a aquest punt, és necessari validar si a les pestanyes de dades de visualitzacions i clics (en funció de la xarxa social) hi ha totes les combinacions possibles. Excel mostra que per cada xarxa social tenim 371.600 registres, aquesta dada no conté totes les combinacions possibles de valors. El càlcul de combinacions possibles seria el següent:

$$365\ dies\ x\ 16\ zones\ x\ 8\ productes\ x\ 4\ rangs\ edat\ x\ 2\ gèneres\ x\ 8\ aficions = 2.990.080\ registres$$

Només tenim informació d'aproximadament un 12,5% de les possibles combinacions. Aquesta mancança de combinacions no suposa cap problema, ja que en el disseny del DW els diferents conceptes estaran en dimensions i a la taula de fets només hi apareixeran a les combinacions reals.

En entorns de treball real, no és comú que l'origen de dades sigui un fitxer Excel, sinó que generalment són sistemes transaccionals de diferent naturalesa (heterogenis). Per aquest motiu es realitzarà una transformació simple de les dades, passant-les del fitxer Excel a una BD en MySQL. Tenir les dades en una BD, permetrà implementar de forma més còmode els processos ETL de càrrega de dades en el DW.

El mapeig de la informació entre les dades del fitxer Excel i la BD és directe, de manera que cada pestanya del fitxer Excel es correspon a una taula en la BD i cada columna d'una pestanya es correspon a un atribut en la taula corresponent. A continuació es mostra una taula on es pot observar la correspondència entre les dades del fitxer Excel y la BD:

Excel		BD			
Pestanya	Columna	Taula	Atribut	Tipus de dades	Mida
Products	Product	Products	Product	VARCHAR	50
Products	Family	Products	Family	VARCHAR	50
Zones	Zone	Zones	Zone	VARCHAR	50
Zones	City	Zones	City	VARCHAR	50
Zones	ZipCode	Zones	ZipCode	VARCHAR	50
Instagram	Date	Instagram	Date	VARCHAR	10
Instagram	ZipCode	Instagram	ZipCode	VARCHAR	50
Instagram	Product	Instagram	Product	VARCHAR	50
Instagram	Age	Instagram	Age	VARCHAR	5
Instagram	Gender	Instagram	Gender	VARCHAR	10
Instagram	Likes	Instagram	Likes	VARCHAR	50
Instagram	Prints	Instagram	Prints	INTEGER	

Instagram	Hits	Instagram	Hits	INTEGER	
Facebook	Date	Facebook	Date	VARCHAR	10
Facebook	ZipCode	Facebook	ZipCode	VARCHAR	50
Facebook	Product	Facebook	Product	VARCHAR	50
Facebook	Age	Facebook	Age	VARCHAR	5
Facebook	Gender	Facebook	Gender	VARCHAR	10
Facebook	Likes	Facebook	Likes	VARCHAR	50
Facebook	Prints	Facebook	Prints	INTEGER	
Facebook	Hits	Facebook	Hits	INTEGER	
Youtube	Date	Youtube	Date	VARCHAR	10
Youtube	ZipCode	Youtube	ZipCode	VARCHAR	50
Youtube	Product	Youtube	Product	VARCHAR	50
Youtube	Age	Youtube	Age	VARCHAR	5
Youtube	Gender	Youtube	Gender	VARCHAR	10
Youtube	Likes	Youtube	Likes	VARCHAR	50
Youtube	Prints	Youtube	Prints	INTEGER	
Youtube	Hits	Youtube	Hits	INTEGER	
Twitter	Date	Twitter	Date	VARCHAR	10
Twitter	ZipCode	Twitter	ZipCode	VARCHAR	50
Twitter	Product	Twitter	Product	VARCHAR	50
Twitter	Age	Twitter	Age	VARCHAR	5
Twitter	Gender	Twitter	Gender	VARCHAR	10
Twitter	Likes	Twitter	Likes	VARCHAR	50
Twitter	Prints	Twitter	Prints	INTEGER	
Twitter	Hits	Twitter	Hits	INTEGER	

3.2 Disseny del Data Warehouse

Després de l'anàlisi i comprensió de les dades del problema que es vol resoldre s'inicia el procés de disseny efectiu del magatzem de dades. La selecció del model de DW (o esquema) marcarà el camí a seguir en les fases següents. És per aquest motiu que s'ha d'analitzar quin tipus d'esquema és el més adient per al problema a resoldre.

Els dos tipus de dissenys bàsics de DW són: disseny en estralla i disseny en "floc del neu" (snowflake). Comentarem les principals característiques de cadascun d'aquests tipus de disseny, per poder realitzar una selecció basada en els requeriments del problema a solucionar. Notar que en el disseny d'un DW no es tenen en compte els esquemes tradicionals de disseny de BD basats en formes normals, ja que el que es pretén es prioritzar el rendiment de les consultes.

El model d'estrella presenta les característiques següents:

- L'estructuració de la informació es fa de manera que tenim una taula de fets central (el que analitzem) i múltiples taules de dimensions relacionades amb la taula de fets. Aquesta topologia recorda a una estrella.
- En general es tindran tantes taules de dimensions com dimensions d'anàlisi participin en la definició del fet estudiat.

- En aquest esquema, només la taula de fets manté *joins* cap a les altres taules en forma de *foregin keys*.
- Les taules de dimensions no estan normalitzades. Aquesta característica fa que es dupliquin dades de forma innecessària.
- Facilitat de comprensió (pels implicats en el projecte de BI) i simplicitat de disseny.
- Genera complicacions en l'actualització de dades "mestres" (per exemple els noms dels productes).
- Proporciona un gran rendiment.

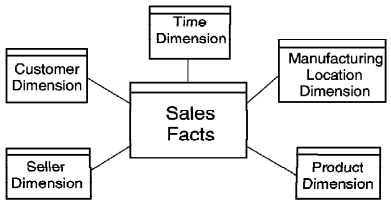


Figura 4: Exemple d'esquema en estrella

L'esquema de floc de neu presenta les característiques següents:

- Aquest esquema deriva de l'esquema en estrella. La principal diferencia és que les taules de dimensions es normalitzen (3NF) en altres taules.
- Aquesta normalització fa que la taula de fets ja no sigui la única a relacionar-se amb altres taules, ara les dimensions també es relacionen amb altres taules.
- Al normalitzar les taules de dimensions s'aconsegueix un estalvi d'espai al no duplicar la informació.
- La informació està més estructurada, però tècnicament és més complex de dissenyar i de comprendre.
- Es senzill actualitzar dades "mestres" (per exemple dades de productes).
- Rendiment sensiblement pitjor que l'esquema d'estrella.

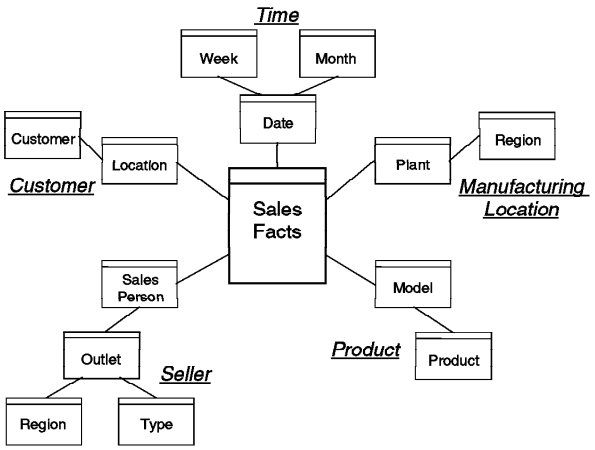


Figura 5: Exemple d'esquema en floc de neu

L'esquema triat per donar solució al cas d'estudi és un disseny dimensional en forma d'estrella. Els motius en els que es basa aquesta decisió són els següents:

- El sistema a modelar és simple. Afegir complexitat innecessàriament aplicant un model en floc de neu no és aconsellable. Els esquemes en floc de neu es reserven per sistemes molt grans i complexos.
- No hi ha problemes d'espai i per tant, tenir informació no normalitzada a les taules no representa un problema des d'aquest punt de vista.
- No hi ha un requeriment clar que ens obligui a actualitzar dades "mestres" de forma continuada.
- En general els sistemes BI prioritzen la velocitat en l'accés a dades, i el nostre cas no és una excepció.
- La simplicitat de disseny afavoreix la comprensió del disseny del DW per part d'altres implicats en el projecte de BI.

Un cop triat l'esquema de DW a utilitzar, s'ha d'iniciar el disseny. Com que el model triat és en estrella, podem seguir les pautes proporcionades a [4]. Aquestes pautes estableixen la base del disseny dimensional de DW i es concreten en quatre etapes de disseny:

1. Definir el procés de negoci que s'està modelant.
2. Definir la granularitat que es tindrà en compte.
3. Identificar les dimensions.
4. Identificar els fets.

A continuació es desenvolupa cadascuna de les etapes.

3.2.1 Identificació del procés de negoci

Identifiquem com a processos de negoci, aquelles activitats operacionals que desenvolupa una empresa o organització. Aquestes activitats operacionals porten associades dades, que interpretades correctament, generen mètriques. Aquestes mètriques serveixen per a l'anàlisi de les operacions i per a la presa de decisions. Identificar correctament el procés de negoci que es vol modelar és indispensable per, posteriorment, definir la granularitat i identificar dimensions i fets.

En el nostre cas d'estudi podem assimilar el problema a resoldre a un procés de negoci específic d'un departament de màrqueting d'una empresa. El departament té la necessitat d'avaluar de forma objectiva l'efectivitat de les campanyes publicitàries; i en base a la informació obtinguda prendre decisions. Aquesta focalització departamental ens duria al disseny d'un Data Mart específic del departament; que formaria part del DW global de la companyia. Tot i això, assimilem que s'està dissenyat un DW, sense tenir en compte factors externs.

Finalment, podríem formular el procés de negoci que ens ocupa com a:

"L'estudi de l'efectivitat de les campanyes de publicitat en mitjans digitals, concretament a xarxes socials, a partir de l'indicador bàsic CTR"

3.2.2 Definició de la granularitat

Establir o definir de forma correcta la granularitat amb la que es representaran les dades és molt important. Aquesta definició de granularitat ens marcarà de forma inequívoca que és el que representa un registre concret a la taula de fets. Abans de definir les dimensions i els fets, s'ha de definir la granularitat perquè tant les dimensions com els fets han de ser consistents amb la granularitat.

Establir una granularitat poc especifica penalitzarà als usuaris, ja que no podran obtenir un nivell de detall de les dades alt. Un dels requeriments del nostre problema ens indica que hem de ser capaços d'oferir informació als usuaris a nivell de dia, i per tant aquesta serà la nostra granularitat. Després amb operacions d'acumulació es podran obtenir dades per mesos o trimestres, però els registres de la taula de fets han de contenir la informació a nivell de dia.

Un cop s'ha establert la granularitat podem enunciar quin serà el contingut de la taula de fets:

"Cada registre de la taula de fets contindrà la informació d'un snapshot diari agrupat per xarxa social, producte, localització i gustos de l'audiència, indicant les publicacions i clics als anuncis."

3.2.3 Identificació de les dimensions

Les dimensions proporcionen el context necessari per entendre el *"qui, què, com, quan, perquè i on"* del fets. Les dimensions contenen els atributs descriptius en un sistema de BI que ens permeten entendre, filtrar i agrupar els fets. També es poden entendre les dimensions com les coses "que importen" del nostre problema. El procés de identificació de dimensions es realitza a través de l'anàlisi de les dades del problema. En el nostre cas d'estudi trobem un total de 7 dimensions:

- *Product dimension*: Aquesta dimensió fa referència a la informació dels productes que s'anuncien.
- *Location dimension*: Aquesta dimensió fa referència a la informació relacionada a la localització geogràfica de les persones (audiència) a les que se lis mostren els anuncis publicitaris.
- *Network dimension*: Aquesta dimensió fa referència a la informació de les diferents xarxes social sobre les que es mostren els anuncis publicitaris. Contindrà la relació de xarxes socials on es mostren els anuncis.
- *Date dimension*: Dimensió temporal que conté la informació de les dates. És una dimensió bàsica en els projectes de BI, ja que ens permet analitzar les dades i el comportament dels fet al llarg del temps. Per exemple, permet comparar l'efectivitat dels anuncis entre mesos, o el primer dia de cada mes. En el cas que ens ocupa només tindrem informació d'un any.
- *Age dimension*: Aquesta dimensió contindrà la relació d'intervals d'edat dels usuaris als que se'ls hi ha mostrat l'anunci.
- *Gender dimension*: Dimensió que contindrà el gènere (sexe) dels usuaris als que se'ls hi ha mostrat l'anunci.
- *Likes dimension*: Aquesta dimensió contindrà la informació de les aficions dels usuaris als que se'ls hi ha mostrat l'anunci.

Aquesta descomposició de dimensions ens permetrà afrontar de forma correcta les operacions de "*Drill Down*" i "*Roll Up*" quan es consulti la informació a traves de cubs OLAP.

Com es pot observar ens proporcionen el context complert sobre el que esdevé un fet de la nostra taula de fets.

3.2.4 Identificació dels fets

Els fets són les mesures que s'obtenen de l'observació d'un procés de negoci i generalment es corresponen amb valor numèrics. A la taula de fets només podem trobar fets que es corresponguin amb la definició de granularitat, i que tinguin sentit per al procés de negoci que s'analitza.

En el nostre cas d'estudi tindrem un únic fet, i per tant una única taula de fets. El fet a analitzar es correspon amb el rendiment dels anuncis segons plataforma que els publica. La informació que ens permet tenir el valor del referència del rendiment de l'anunci és el CTR, que es calcula a partir dels valors d'impressions de l'anunci (vegades que es mostra als usuaris) i de clics (vegades que els usuaris han reaccionat a l'anunci i han clicat en ell). El CTR és una mesura percentual de rendiment que es calcula de la forma següent:

$$CTR = \frac{Clics}{Impressions}\ x\ 100$$

A la taula de fets es tindran en compte les mesures d'impressions i clics per cada dia agrupant per xarxa social, localització, producte, edat, gènere i aficions. Aquesta informació es representa de forma atòmica, és a dir un registre per dia i combinació registrada dels paràmetres; i no totes les combinacions possibles de paràmetres tindran un registre a la taula de fets ja que només s'inclouran les combinacions que han generat dades reals.

3.2.5 Consideracions de disseny

Un cop s'ha triat el model i definides les dimensions i fets, és convenient repassar tres decisions més de disseny del DW. El primer aspecte a comentar resideix en el tipus de claus primàries que s'empraran a les taules de dimensions. Aquestes claus seran claus subrogades, és a dir que no es farà servir la clau natural de la taula. Enlloc es farà servir un atribut desvinculat de tipus enter i que serà autonumèric. D'aquesta manera aconseguim desvincular el DW de les possibles claus dels orígens de dades.

El segon aspecte remarcable és que a la taula de fets, no s'inclourà un atribut que contingui el càlcul del CTR. Com s'ha comentat al punt 3.2.4, el CTR és una mesura de tipus percentatge (*non-additive fact*). Les mesures de tipus percentatge no són sumables a través de les dimensions, i per tant ens trobaríem amb problemes a l'hora de consultar la informació. La recomanació és no mantenir aquest valor calculat a la taula de fets, i calcular-lo quan es realitzin les operacions de "*Drill Down*" i "*Roll Up*" en els informes (cubs OLAP)

El darrer aspecte a comentar resideix en la identificació dels atributs de les taules de dimensions del DW amb la seva corresponent SCD (*Slowly Changing Dimension*). En les modelitzacions de DW ens trobem amb que la informació de les dimensions, generalment canvia (s'afegeix informació, es modifica, s'elimina, etc.). La forma en la que es gestionen aquests canvis està relacionada amb els tipus de SCD. En el nostre DW assumim, que no es canviaran les dades (són estàtiques o persistents) i per tant les identificarem amb un SCD de tipus 0. Per una referencia més completa dels tipus de SCD es pot consultar [6] a la bibliografia.

3.2.6 Modelització del DW

A continuació es mostren dues vistes del disseny conceptual del DW. En la primera es mostra el disseny en alt nivell (sense detalls de les taules de dimensions), mentre que el segon mostra detalladament el disseny de cadascuna de les taules (tant de fets com de dimensions).

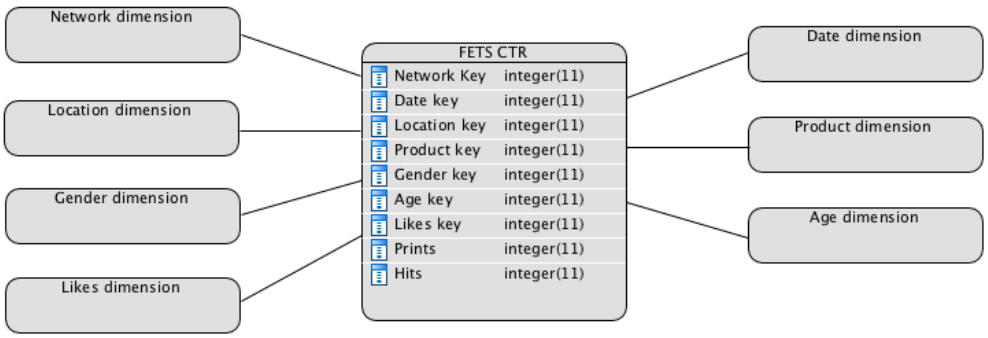


Figura 6: Disseny simplificat del DW

A continuació el disseny detallat del DW:

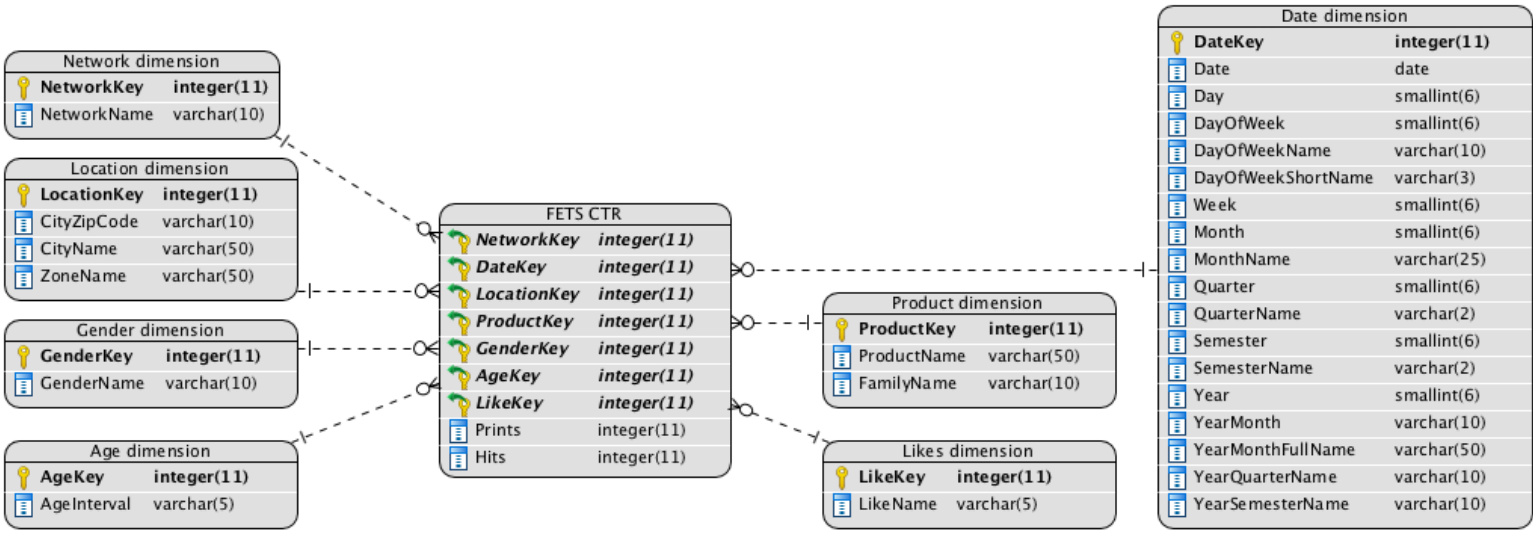


Figura 7: Disseny detallat del DW

3.2.7 Detall de les taules del DW

En els projectes de BI (com en la resta de projectes tecnològics de Sistemes d'Informació), una de les parts fonamentals en el desenvolupament del projecte és la documentació. Per això és necessari documentar de forma correcta, i el més detalladament possible, l'estructura de les taules del DW. És per això que a continuació s'inclou la informació detallada de totes les taules que formen part del DW (taula de fets i taules de dimensions).

El model de document que s'ha seguit fa referència a [5], i mostra el disseny de les taules del DW i la seva relació amb l'origen de dades:

- **Taula "Network"**

Nom	NetworkDim
Tipus	Dimensió
Nom mostrat	Network
Descripció	La dimensió "Network" conté la informaciób àsica de les diferents xarxes socials que mostren anuncis de productes
Mida	4 registres

Destí						Origen			
Camp	Descripció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
NetworkKey	Clau primaria subrogada	SMALLINT		1,2,3,...	0				
NetworkName	Nom de la xarxa social	VARCHAR	10	Youtube	0	DataPubli	Youtube, Instagram, Facebook,Twitter	N/A	N/A

La informació que conté la taula "Network" s'obté de forma general ja que a les dades d'origen, aquesta informació està en forma de "nom" de taula, és a dir que per cada taula de dades de xarxa social, crearem un registre a la taula "Network". No hi ha un mapeig directe entre l'origen de dades i la taula al DW.

- **Taula "Location"**

Nom	LocationDim
Tipus	Dimensió
Nom mostrat	Location
Descripció	La dimensió "Location" conté la informació bàsica de les diferents localitzacions registrades al sistema
Mida	16 registres

						Origen			
Camp	Descripció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
LocationKey	Clau primaria subrogada	SMALLINT		1,2,3,...	0				
CityZipCode	Codi postal de la localització	VARCHAR	10	48003	0	DataPubli	Zones	ZipCode	VARCHAR
CityName	Ciutat de la localització	VARCHAR	50	Bilbao	0	DataPubli	Zones	City	VARCHAR
ZoneName	Zona geogràfica de la localització	VARCHAR	50	North	0	DataPubli	Zones	Zone	VARCHAR

Les dades de la taula de dimensió "Location" s'obtenen de la informació de la taula "Zones" original.

- **Taula “Product”**

Nom	ProductDim
Tipus	Dimensió
Nom mostrat	Product
Descripció	Taula de productes-families que estan registrats al sistema
Mida	8 registres

Destí						Origen			
Camp	Descipció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
ProductKey	Calu primària subrrogada	SMALLINT		1,2,3,...	0				
PrductName	Nom del producte	VARCHAR	50	Mobile Phone	0	DataPubli	Product	Product	VARCHAR
FamilyName	Nom de la familia del producte	VARCHAR	50	Electronics	0	DataPubli	Product	Family	VARCHAR

Les dades de la taula de dimensió “Product” s’obtenen de la informació de la taula “Product” original.

- **Taula “Gender”**

Nom	GenderDim
Tipus	Dimensió
Nom mostrat	Gender
Descripció	Taula que conté la informació bàsica del sexe dels usuaris
Mida	2 registres

Destí						Origen			
Camp	Descipció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
GenderKey	Clau primària subrrogada	SMALLINT		1,2,3,...	0				
GenderName	Nom del genere (sexe)	VARCHAR	10	Male	0	DataPubli	Youtube	Gender	VARCHAR

Aquesta taula s’omple amb la informació de gèneres continguda a la taula Youtube. Es podria obtenir de qualssevol altra taula de xarxa social (Instagram, Twitter o Facebook), ja que la informació relativa a generes és la mateixa a les 4 taules.

- **Taula “Age”**

Nom	AgeDim
Tipus	Dimensió
Nom mostrat	Age
Descripció	Taula que conté els rangs d'edat dels usuaris del sistema.
Mida	4 registres

Destí						Origen			
Camp	Descipció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
AgeKey	Clau primària subrrogada	SMALLINT		1,2,3,...	0				
AgeInterval	Rang d'edat	VARCHAR	5	18-30	0	DataPubli	Youtube	Age	VARCHAR

Aquesta taula s’omple amb la informació de rangs d’edat continguda a la taula Youtube. Es podria obtenir de qualssevol altra taula de xarxa social (Instagram, Twitter o Facebook), ja que la informació relativa a rangs d’edat és la mateixa a les 4 taules.

- **Taula “Likes”**

Nom	LikesDim
Tipus	Dimensió
Nom mostrat	Likes
Descripció	Taula que conté la informació bàsica dels gustos o aficions dels usuaris
Mida	8 registres

Destí						Origen			
Camp	Descipció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
LikesKey	Clau primària subrrogada	SMALLINT		1,2,3,...	0				
LikeName	Nom de l'afició o "gust"	VARCHAR	10	Male	0	DataPubli	Youtube	Likes	VARCHAR

Aquesta taula s’omple amb la informació d’aficions continguda a la taula Youtube. Es podria obtenir de qualssevol altra taula de xarxa social (Instagram, Twitter o Facebook), ja que la informació relativa a aficions es la mateixa a les 4 taules.

- **Taula “Date”**

Nom	DateDim
Tipus	Dimensió
Nom mostrat	Date
Descripció	Taula que conté eldesenvolupament de dates d'un any (datesde les dades de l'any 2017)
Mida	365 registres

						Origen			
Camp	Descipció	Tipus	Mida	Valors d'exemple	Tipus SCD	Sistema Origen	Taula origen	Camp origen	Tipus
DateKey	Clau primèria subrogada	SMALLINT		1	0				
Date	Data en format dd/MM/AAAA	DATE		01/01/2017	0	DataPubli	Youtube	Date	DATE
Day	Dia al que es correspon el camp 'Date'	SMALLINT		31	0				
DayOfWeek	Ordinal de dia de la setmana que es correspon amb el camp 'Date'	VARCHAR		7	0				
DayOfWeekName	Nom del dia de la setmana que es correspon a 'DayOfWeek'	VARCHAR	10	Monday	0				
DayOfWeekShortName	Abreviatura del camp 'DayOfWeekName'	VARCHAR	3	Monday	0				
Week	Setmana a la que correspon el camp 'Date'	SMALLINT		1	0				
Month	Mes al que es correspon el camp 'Date'	SMALLINT		1	0				
MonthName	Nom del mes al que es correspon el camp 'Date'	VARCHAR	25	January	0				
Quarter	Trimestre al que es correspon el camp 'Date'	SMALLINT		1	0				
QuarterName	Nom del trimestre al que es correspon el camp 'Date'	VARCHAR	2	Q1	0				
Semester	Semestre al que es correspon el camp 'Date'	SMALLINT		1	0				
SemesterName	Nom del semestre al que es correspon el camp 'Date'	VARCHAR	2	S1	0				
Year	Any del camp 'Date'	SMALLINT		2017	0				
YearMonth	Any i mes al que es correspon el camp 'Date'	VARCHAR	10	01/2017	0				
YearMonthFullName	Nom complert de l'any i mes del camp 'Date'	VARCHAR	50	2017 January	0				
YearQuarterName	Nom complert de l'any i trimestre del camp 'Date'	VARCHAR	10	2017-Q1	0				
YearSemesterName	Nom complert de l'any i semestres del camp 'Date'	VARCHAR	10	2017-S1	0				

Aquesta taula s’omple amb les dates contingudes a la taula Youtube. Es podria obtenir de qualssevol altra taula de xarxa social (Instagram, Twitter o Facebook), ja que la informació relativa a dates és la mateixa a les totes les taules de dades de xarxes socials. La resta de camps de la taula es calculen automàticament en base a la informació continguda al camp “Date”.

- **Taula “CTR”**

Nom	CTRFact
Tipus	Fets
Nom mostrat	CTR
Descripció	Taula que conté la relació sencera de fets de visualització i clics d'anuncis agrupada per data, xarxa social, localització, producte, i dades d'usuari (sexe, edat i aficions)
Mida	1.486.400 registres

Destí					Origen			
Camp	Descripció	Tipus	Mida	Valors d'exemple	Sistema Origen	Taula origen	Camp origen	Tipus
NetworkKey	Clau de xarxa social	SMALLINT		1				
DateKey	Clau de data	SMALLINT		128	DataPubli	Youtube, Instagram, Facebook,Twitter		
LocationKey	Clau de loaclització	SMALLINT		7	DataPubli	Youtube, Instagram, Facebook,Twitter		
ProductKey	Clau de producte	SMALLINT		5	DataPubli	Youtube, Instagram, Facebook,Twitter		
GenderKey	Clau de gènere (sexe)	SMALLINT		2	DataPubli	Youtube, Instagram, Facebook,Twitter		
AgeKey	Clau d'edat	SMALLINT		4	DataPubli	Youtube, Instagram, Facebook,Twitter		
LikeKey	Clau d'aficions	SMALLINT		6	DataPubli	Youtube, Instagram, Facebook,Twitter		
Prints	Número d'impressions de l'anunci segons l'agrupació de dades del registre	SMALLINT		785	DataPubli	Youtube, Instagram, Facebook,Twitter	Prints	VARCHAR
Hits	Clics a l'anunci segons l'agrupació de dades del regitre	SMALLINT		44	DataPubli	Youtube, Instagram, Facebook,Twitter	Hits	VARCHAR

La taula CTR és la taula de fets. Cada fet conté la informació d’impressions i clics d’un anunci agrupat per xarxa social, data, localització, producte, gènere, edat i aficions. L’origen de dades serà la unió de la informació continguda a totes les taules de dades d’impressions i clics de cadascuna de les xarxes socials.

La clau primària de la taula de fets està composada per: NetworkKey, DateKey, LocationKey, ProductKey, GenderKey, AgeKey, LikeKey.

NetwokKey és clau forana a la taula de dimensió Network.

DateKey és clau forana a la taula de dimensió Date

LocationKey és clau forana a la taula de dimensió Location

ProductKey és clau forana a la taula de dimensió Product

GenderKey és clau forana a la taula de dimensió Gender

AgeKey és clau forana a la taula de dimensió Age

LikeKey és clau forana a la taula de dimensió Likes

4. Modelització i implementació dels processos ETL.

4.1 Introducció a la modelització de processos ETL.

Tot magatzem de dades ha de contenir la informació necessària per dotar al sistema de BI de la funcionalitat requerida. Aquesta informació pot arribar des d'orígens heterogenis, barrejant informació de Bases de Dades de diferents proveïdors (per exemple SQL Server i PostgreeSQL) juntament amb fitxers de dades en Excel o simples fitxers de text. Aquesta informació s'ha d'obtenir, transformar-la per adequar-la a les regles del negoci que han de ser suportades per el DW i finalment s'han d'injectar al DW per tal de poder ser explotades. En aquest capítol es donarà una visió genèrica dels processos ETL i del perquè de la necessitat de modelar aquests processos, per a continuació implementar de forma efectiva els processos ETL requerits al nostre problema emprant les eines que posa a disposició la plataforma BI seleccionada (Pentaho).

La modelització i implementació dels processos ETL és una de les parts més costoses dels projectes BI donada la complexitat d'obtenció i transformació de les dades. Es calcula, que en funció de la complexitat del projecte, aquesta etapa pot durar entre un 30% i un 80% del temps total del projecte [19]. Els processos ETL estan fortament lligats a requeriments tècnics de rendiment, integritat, seguretat i confiabilitat; a més un cop implementats hi ha d'haver necessàriament una validació de la qualitat de les dades injectades al DW. Aquesta validació ha de servir per assegurar que les dades del DW són correctes, no entren en conflicte entre elles ni amb les regles del negoci, i en cas d'haver calculat algunes dades, aquests càlculs han de ser correctes. Qualssevol error en les dades injectades pot portar a conclusions equivocades sobre els processos de negoci en els que s'aplicarà el BI, amb el consegüent perjudici econòmic per a l'empresa. Per tots els motius esmentats, és d'especial interès que els processos ETL estiguin correctament identificats i dissenyats.

Aquesta modelització dels processos ETL (o definició formal), es fa seguint la notació bàsica d'UML però centrant-se en la vinculació d'atributs entre els orígens de dades i el destí (correspondència de les dades).

A continuació s'esposa la notació bàsica per a la modelització de processos ETL i el significat de cadascun dels elements:

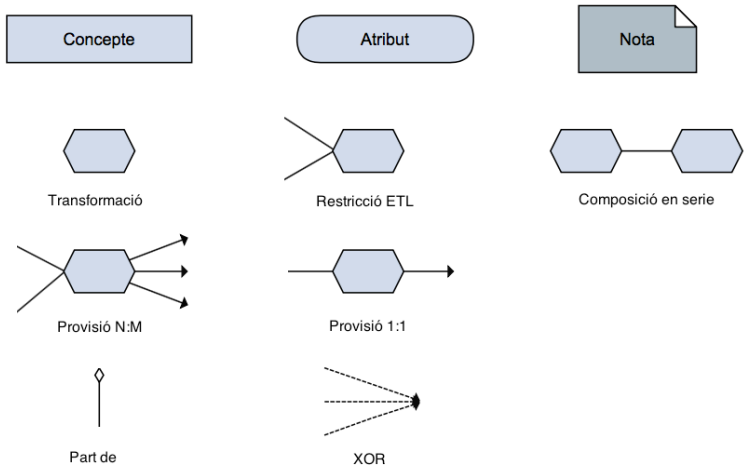


Figura 8: Elements per modelar processos ETL

- *Concepte:* representen les entitats d'informació. Poden ser entitats de BD de l'origen de dades, taules de dimensions o de fets del DW, fitxers d'informació, etc.
- *Atribut:* representa la unitat d'informació mínima d'una entitat.
- *Part de:* Estableix la vinculació entre el concepte i els seus atributs, ja que el concepte està format per atributs.
- *Notes:* Comentaris addicionals per aclarir certs aspectes que la notació no defineix. Per exemple: la funció f() de transformació, converteix un import de USD a EUR.
- *Transformació:* representen tasques simples que es realitzen dins del marc de treball ETL. Per exemple trobar els valors màxim i mínim d'un conjunt de dades.
- *Restricció ETL:* representen que les dades d'un conjunt han de complir necessàriament certs requeriments.
- *Composició en sèrie:* representa la necessitat d'encadenar diverses etapes de transformacions en una provisió 1:1.
- *Provisió (1:1 i N:M):* Representen el mapeig d'atributs entre un origen i un destí.
- *XOR:* representa la situació en la que diversos orígens de dades (fitxers, taules de BD, etc.) contenen la informació a injectar a la taula destí del DW.

Es pot obtenir informació informació més detallada dels elements de notació a la referència bibliogràfica [19].

4.2 Modelització dels processos ETL

A continuació es mostren els models dels processos de transformació aplicats al problema que s'està resolent.

4.2.1 Modelització del procés generació de la dimensió *Network*

La transformació per generar la informació de la dimensió *Network* és senzilla. Es tracta d'una vinculació directa dels noms de les taules que contenen les dades d'impressions i clics cap a la taula de dimensió. Aquests noms de taula es corresponen amb les diferents xarxes socials a considerar. La taula de dimensió té una clau primària subrogada que es calcula de forma automàtica (auto-increment de MySQL).

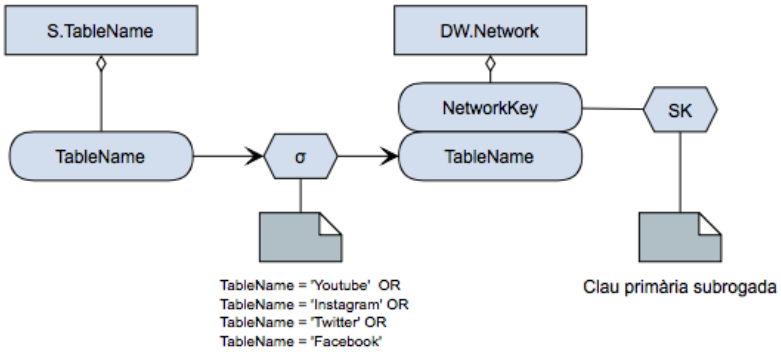


Figura 9: Modelització del procés de generació de la dimensió *Network*

4.2.2 Modelització del procés de generació de la dimensió *Product*

La transformació per generar la informació de la dimensió *Product*, és una vinculació directa de la informació de la taula d'origen (Products) cap a la taula de destí del DW. La taula de dimensió té una clau primària subrogada que es calcula de forma automàtica (auto-increment de MySQL).

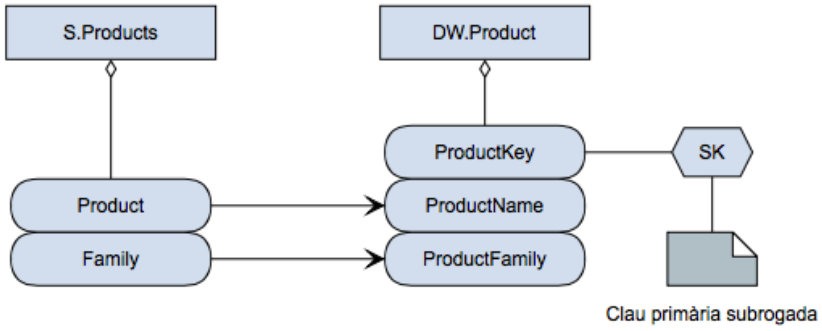


Figura 10: Modelització del procés de generació de la dimensió *Product*

4.2.3 Modelització del procés de generació de la dimensió *Location*

La transformació per generar la informació de la taula *Location*, és una vinculació directa de la informació de la taula d'origen (Zones) cap a la taula de destí del DW. La taula de dimensió té una clau primària subrogada que es calcula de forma automàtica (auto-increment de MySQL).

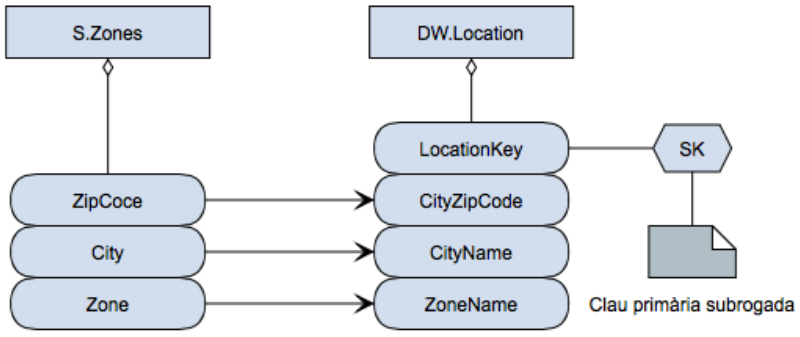


Figura 11: Modelització del procés de generació de la dimensió *Location*

4.2.4 Modelització del procés de generació de la dimensió *Date*

A continuació es mostra el modelat ETL de la dimensió *Date*, L'origen de dades de la transformació, és el camp data de les taules d'impressions i clics de cadascuna de les xarxes

socials. S'obtenen totes les dades de dates i es transformen de cadena de text a data. A continuació cerquem la data màxima i mínima per establir els intervals de data inicial i final sobre els que generarem les dades de la dimensió. Un cop tenim aquesta informació apliquem les transformacions sobre les dates (obtenció del dies, mes, any, trimestre, dia del mes, etc.) per tal d'obtenir la resta d'informació necessària per omplir la taula de destí del DW La taula de dimensió té una clau primària subrogada que es calcula de forma automàtica (auto-increment de MySQL).

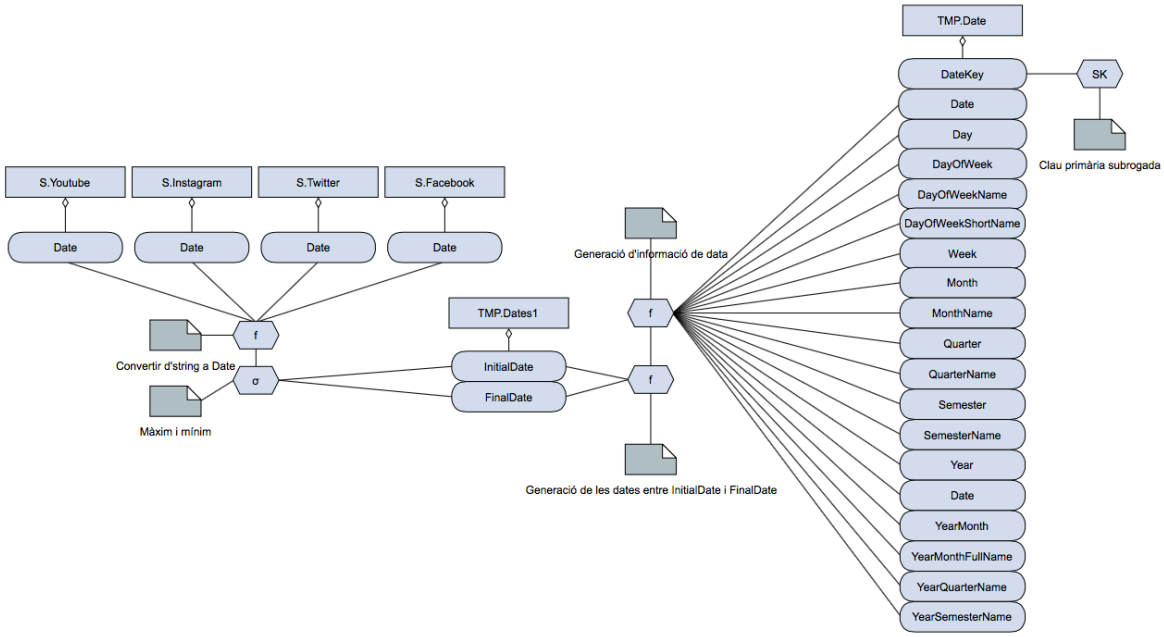


Figura 12: Modelització del procés de generació de la dimensió *Date*

4.2.5 Modelització del procés de generació de les dimensions *Age, Gender* i *Likes*

La modelització dels processos de generació de les transformacions *Age, Gender* i *Likes* segueixen el mateix esquema conceptual. S'obtenen les dades del camp a transformar de les taules de dades d'impressions i clics. S'obtenen els valors únics de totes les taules del paràmetre que s'està tractant (*Age*, *Gender* o *Likes*). Els valors obtinguts es vinculen a la taula de dimensió corresponent. Les taules de dimensió tenen una clau primària subrogada que es calcula de automàticament (auto-increment de MySQL).

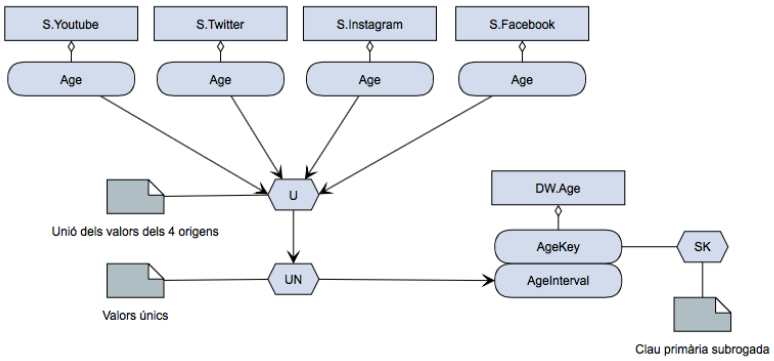


Figura 13: Modelització del procés de generació de la dimensió *Age*

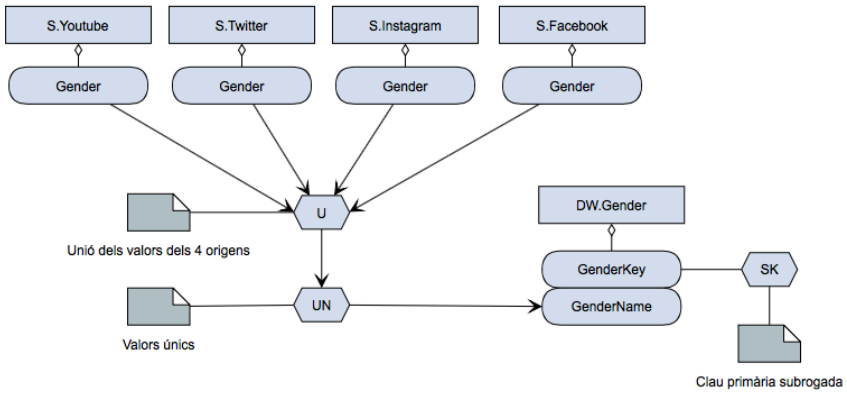


Figura 14: Modelització del procés de generació de la dimensió *Gender*

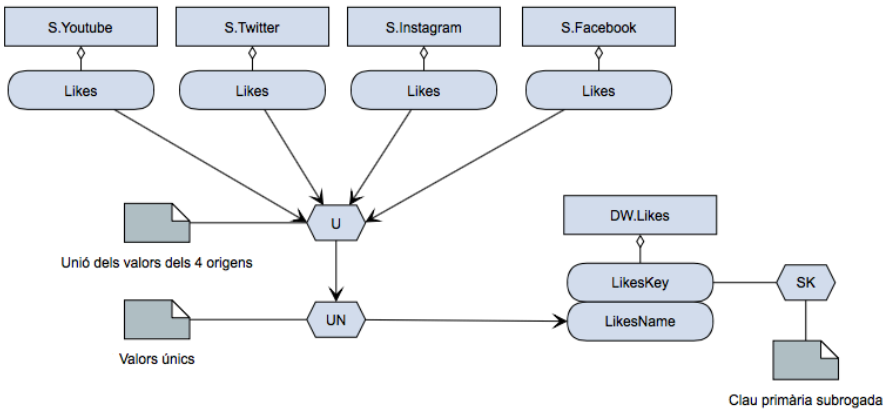


Figura 15: Modelització del procés de generació de la dimensió *Likes*

4.2.6 Modelització del procés de generació de la taula de fets *CTR*

La generació de la informació de taula de fets és la més complexa a nivell de detall de totes les transformacions que es realitzen. Conceptualment la transformació és senzilla: S'uneixen totes les dades de les taules d'impressions i clics i s'envien a la taula de fets. A continuació es pot veure el model general de la transformació:

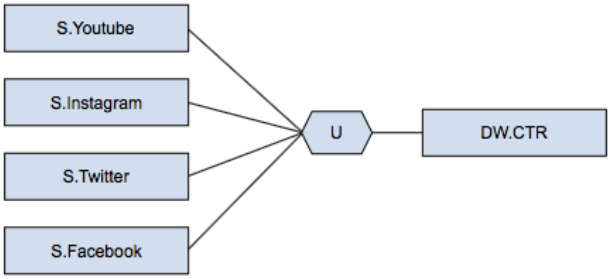
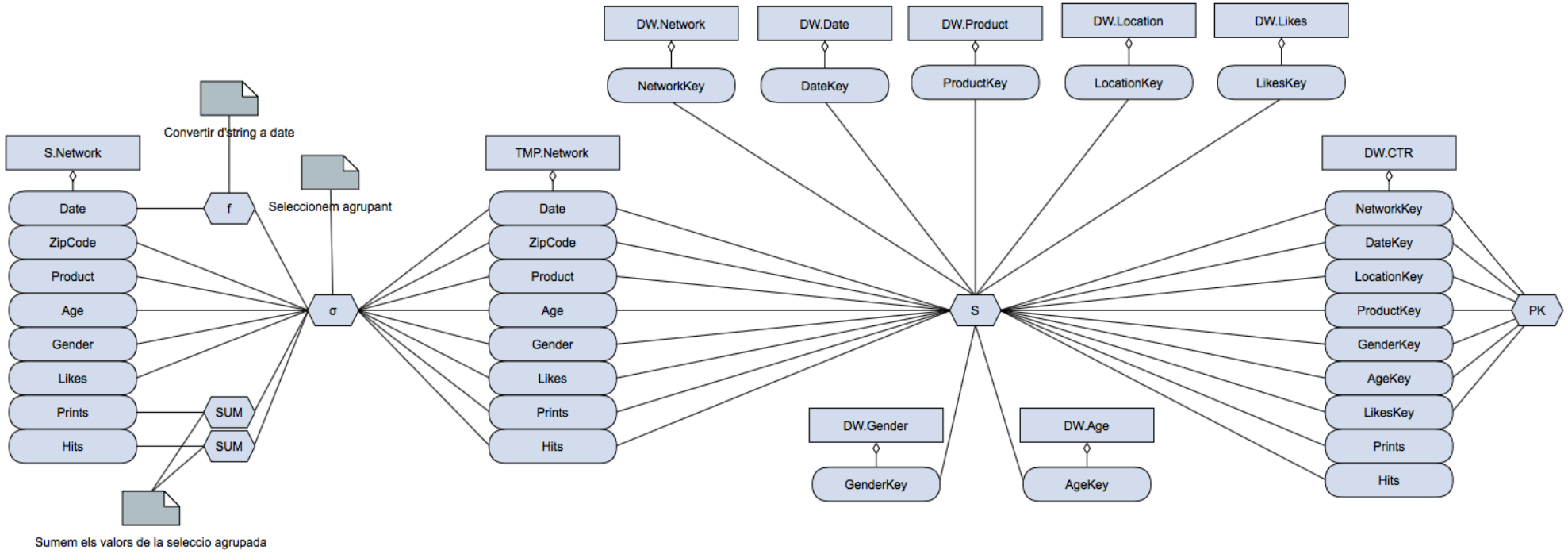


Figura 16: Modelització del procés general de generació de la taula de fets CTR

A continuació es presenta el detall de les transformacions que s'apliquen al conjunt de dades d'origen per generar la informació de la taula de fets *CTR*.



El diagrama anterior modelitza la transformació per una xarxa social (Youtube per exemple). S'obtenen les dades de la taula, canviant el tipus del camp de data, agrupant pels valors (Date, ZipCode, Product, Age, Gender, Likes) per si hi hagés informació per aquest grup repartida en diversos registres.

Per cada grup es sumen els valors de clics i impressions, obtenint les dades llestes per ser enviades a la taula de destí. Abans de poder-ho fer, s'han d'obtenir tots els valors dels camps que formen part de la PK de la taula destí del DW. És a dir, que per cada valor de data, xarxa social, producte, localització, aficions, gènere i edat hem de trobar el seu valor de clau a les taules de dimensió. Un cop es tenen aquests valors ja es pot omplir la taula de destí del DW.

4.3 Implementació dels processos ETL amb Pentaho.

Per a la implementació dels processos ETL a la plataforma BI Pentaho, s'utilitza l'eina Pentaho Data Integration (de nom clau Kettle). Pentaho Data Integration (PID) proporciona als usuaris un ampli conjunt d'eines que permeten realitzar les tasques d'extracció, transformació i càrrega de dades entre sistemes. La interfície gràfica d'usuari (GUI) que permet el disseny de les transformacions en un entorn gràfic i amigable s'anomena Spoon. En general les eines ETL de les plataformes BI són eines que es consideren de metadades perquè treballen a nivell de definició d'accions a realitzar, enlloc de detallar com es realitzen aquestes accions. Quan implementem una transformació a PID ho fem pensant en quines accions s'han de fer deixant de banda el detall de com internament s'han implementat aquestes accions. Aquesta simplicitat de disseny i implementació facilita l'ús de les eines ETL a usuaris que no han de tenir nocions tècniques de programació, només es necessita conèixer l'origen de les dades, quines operacions s'han de realitzar i el destí d'aquestes operacions. A continuació es mostren les implementacions de cadascuna de les transformacions modelades al punt anterior.

4.3.1 Treball global d'execució de totes les transformacions

Totes les transformacions que s'han d'aplicar al conjunt de dades original, s'agrupen en un treball (*job*), de manera que s'executaran de forma seqüencial una darrere l'altra. Notar que tot i que no hi dependència entre algunes transformacions s'executen en sèrie; això es degut a limitacions de l'entorn (Kettle) que no permet l'execució de transformacions en paral·lel. Amb la implementació del treball global s'obté un únic fitxer que realitzarà totes les transformacions, de manera que, un hipotètic client només hauria d'executar el treball global per omplir el magatzem de dades.

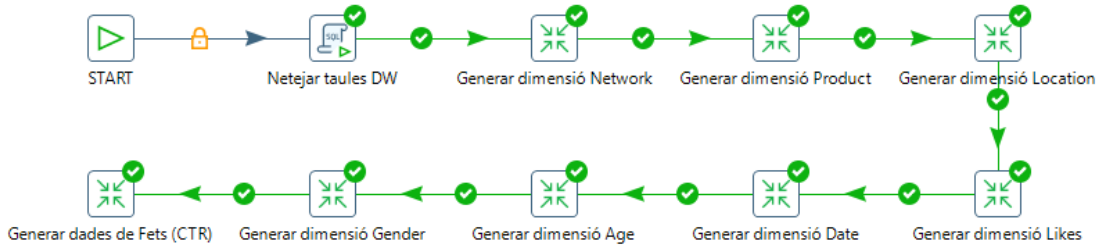


Figura 17: Definició del treball general

El treball és senzill, ja que només encadena transformacions una rere l'altre. Per tal d'assegurar que el treball s'executa sense problemes es realitza una neteja de dades prèvia (pas "Netejar taules DW") i a continuació es realitzen en sèrie totes les transformacions de creació de dimensions, i finalment es genera la taula de fets. Un cop finalitza l'execució del treball (si no hi ha cap error) el DW tindrà totes les dades per tal de poder ser explotades.

4.3.2 Generació de la dimensió *Network*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Network* es mostra a continuació:

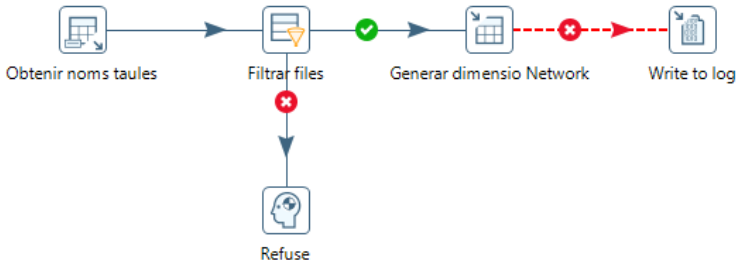


Figura 18: Transformació de generació de la dimensió *Network*

Per a la realització d'aquesta transformació partim del fet que hi ha quatre taules a l'origen de dades que contenen la informació de cadascuna de les xarxes socials a analitzar. El nom d'aquestes taules es correspon amb el nom de les xarxes socials, i és la informació que ens interessa (el nom de la xarxa). La transformació de generació de la dimensió *Network*, es detalla a continuació:

1. S'obtenen tots els noms de taules de l'esquema de BD que s'utilitza com a origen de dades.
2. Les dades obtingudes al pas anterior es filtren de manera que aquells noms de taula que no es corresponen amb una xarxa social queden descartats; mentre que els que concorden amb Youtube, Instagram, Twitter o Facebook superen el filtre i s'empraran en el pas següent
3. Les dades que arriben a aquest pas, s'injecten a la taula de dimensió *Network* del DW (es vinculen els camps i s'injecten les dades).
4. En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Network,* ja té dades que poden ser utilitzades.

4.3.3 Generació de la dimensió *Product*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Product* es mostra a continuació:

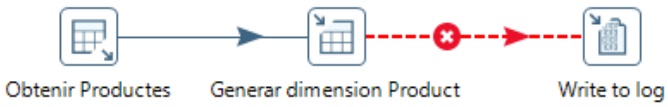


Figura 19: Transformació de generació de la dimensió *Product*

La transformació de generació de la dimensió *Product*, es detalla a continuació:

1. S'obtenen tots els productes definits a la taula "Products" de l'origen de dades.
2. Les dades obtingudes al pas anterior s'injecten a la taula de dimensió *Product* del DW (es vinculen els camps i s'injecten les dades).
3. En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Product,* ja té dades que poden ser utilitzades.

4.3.4 Generació de la dimensió *Location*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Location* es mostra a continuació:

Figura 20: Transformació de generació de la dimensió *Location*

La transformació de generació de la dimensió *Location*, es detalla a continuació:

1. S'obtenen totes les zones definides a la taula "Zones" de l'origen de dades.
2. Les dades obtingudes al pas anterior s'injecten a la taula de dimensió *Location* del DW (es vinculen els camps i s'injecten les dades).
3. En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Location,* ja té dades que poden ser utilitzades.

4.3.5 Generació de la dimensió *Date*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Date* es mostra a continuació:

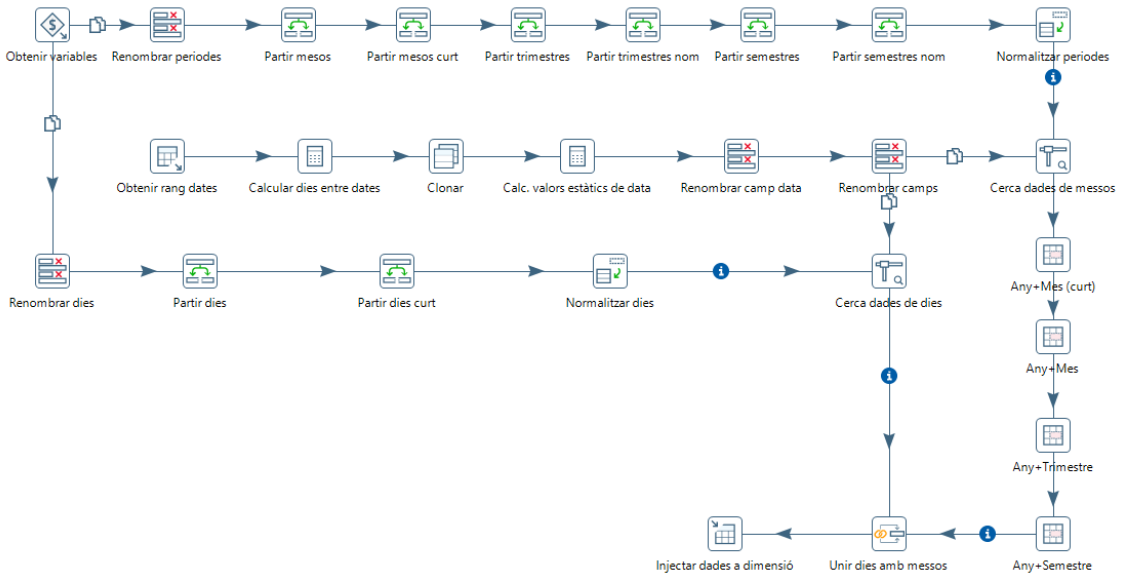


Figura 21: Transformació de generació de la dimensió *Date*

Previ a la definició del primer pas de la transformació s'han de crear una sèrie de paràmetres que s'utilitzaran per realitzar algunes tasques en la transformació. El paràmetres es refereixen al literals de mesos, trimestres, semestres, etc. que seran utilitzats per generar la informació llegible per als usuaris finals. A continuació es mostra la pantalla de definició dels paràmetres de la transformació:

Parameters :

#	Parameter	Default Value	Description
1	DAYS	Sunday,Monday,Tuesday,Wednesday,Thursday,Friday,Saturday	Dies de la setmana
2	DAYS_SHORT	Sun,Mon,Tue,Wed,Thu,Fri,Sat	Nom curt dels dies de la setmana
3	END_DATE		Data final del rang de dates a generar
4	MONTHS	January,February,March,April,May,June,July,August,September,October,November,December	Mesos de l'any
5	MONTHS_SHORT	Jan,Feb,Mar,Apr,May,Jun,Jul,Aug,Sep,Oct,Nov,Dec	Nom curt dels mesos de l'any
6	QUARTERS	1,1,1,2,2,2,3,3,3,4,4,4	Trimestres de l'any
7	QUARTERS_NAME	Q1,Q1,Q1,Q2,Q2,Q2,Q3,Q3,Q3,Q4,Q4,Q4	Nom dels trimestres
8	SEMESTERS	1,1,1,1,1,1,2,2,2,2,2,2	Semestres de l'any
9	SEMETERS_NAME	S1,S1,S1,S1,S1,S1,S2,S2,S2,S2,S2,S2	Nom dels semestres
10	START_DATE		Data inicial del rang dse dates a generar

Figura 22: Paràmetres necessaris per a la transformació de la dimensió *Date*

Juntament amb la transformació encarregada de generar la taula de fets, aquesta transformació és la que més passos realitza. Els passos necessaris per a omplir la dimensió *Date* són els següents:

Passos de la fila superior de la transformació (segons Figura 21):

1. El primer pas és obtenir les variables que es necessitaran per a la transformació.
2. Partim la informació del paràmetre *MONTHS*, en dotze camps de tipus text
3. Partim la informació del paràmetre *MONTHS_SHORT*, en set camps de tipus text.
4. Partim la informació del paràmetre *QUARTERS* en dotze camps de tipus text (cada mes tindrà associat el trimestre que li correspon)
5. Partim la informació del paràmetre *QUARTERS_NAME* nom en dotze camps de tipus text (cada mes tindrà associat el nom del trimestre que li correspon)
6. Partim la informació del paràmetre *SEMESTRERS* en dotze camps de tipus text (cada mes tindrà associat el semestre que li correspon)
7. Partim la informació del paràmetre *SEMESTERS_NAME* en dotze camps de tipus text (cada mes tindrà associat el nom del semestre que li correspon).
8. Normalitzem totes les dades anterior de manera que passem de tenir-les en files a columnes.

Passos de la segona fila de la transformació (segons la Figura 21):

1. Obtenim el rang de dates que s'ha de generar en funció de les dades de les taules d'impressions i clics. Aquest pas es fa en funció de la informació emmagatzemada a l'origen de dades enlloc de fixar paràmetres amb les dates. D'aquesta manera la transformació és més versàtil, i si canviessin les dades d'origen per un interval de dates diferent, la transformació seguirà funcionant correctament sobre el nou interval de dates.
2. Es calculen quants dies hi ha entre les dues dates (en el nostre cas un any).
3. Clonem columna per tal de poder calcular el rang de dates en funció dels dies de diferencia entre les dates i un valor incremental que permetrà generar la seqüència correcta de dates en el pas següent.

4. Es calculen camps de l'stream de dades en funció del camp de data (mes, any, dia del mes, etc.).
5. Renombrem el camp de data per tal de donar format correcte a la sortida.
6. Com en el pas anterior renombrem els altres camps.
7. Amb tot l'stream de dades (un any sencer) realitzem les tasques de cerca en els valors de obtinguts del pas 8 de la primera fila superior. D'aquesta manera generem més camps combinant la informació de dates amb els literals que ajudaran a una millor comprensió de les dades (quan aquestes s'analitzin posteriorment).

En la tercera fila de la transformació (segons la Figura 21) es realitza el mateix procediment que en la primera, però en aquest cas per als dies de la setmana. No es pot realitzar al mateix temps que amb als mesos perquè les dimensions dels paràmetres són diferents, dotze per als mesos i set per als dies de la setmana. Els passos realitzats son els següents:

1. Partim la informació del paràmetre *DAYS*, en set camps de tipus text
2. Partim la informació del paràmetre *DAYS_SHORT*, en set camps de tipus text.
3. Normalitzem totes les dades anterior de manera que passem de tenir-les en files a columnes.
4. Amb tot l'stream de dades (un any sencer) obtingut al pas 6 de la segona fila, realitzem les tasques de cerca dels valors de dies. D'aquesta manera generem més camps combinant la informació de dates amb els literals dels dies que ajudaran a una millor comprensió de les dades (quan aquestes s'analitzin posteriorment).

Abans de procedir a la fusió de les dades dels dos streams (el de dies i el de mesos/trimestres/semestres) generem nous camps a l'stream de mesos realitzant les transformacions següents:

1. Concatenació dels camps Any i Mes curt per generar la informació d'un camp de la taula de destí.
2. Concatenació dels camps Any i Mes per generar la informació d'un camp de la taula de destí.
3. Concatenació dels camps Any i Trimestre per generar la informació d'un camp de la taula de destí.
4. Concatenació dels camps Any i Semestre per generar la informació d'un camp de la taula de destí.

Arribats a aquest punt, els dos streams de dades estan preparats per fusionar-se per finalment injectar les dades al DW. Les transformacions que es realitzen són:

5. Unió (per clau de data) l'stream de dades de dies amb l'stream de dades de mesos, per obtenir un únic stream final.
6. En aquest punt l'stream de dades està preparat per ser injectat a la taula de destí. En aquest pas es vinculen els camps i s'injecten les dades.

Un cop hem arriba a aquest punt la taula de dimensió *Date*, té les dades corresponents a l'interval de dates per al que tenim dades d'impressions i clics.

4.3.6 Generació de la dimensió *Age*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Age* es mostra a continuació:

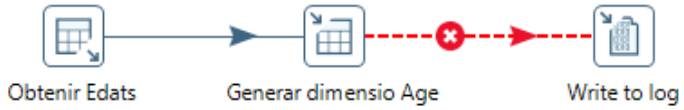


Figura 23:Transformació de generació de la dimensió *Age*

Els passos de transformació per a la obtenció de les dades per a la dimensió *Age*, es detallen a continuació:

1. S'obtenen tots els rangs d'edat diferents de les taules de dades amb una consulta SQL. La consulta uneix els resultats de les taules de dades (Youtube, Instagram, Twitter i Facebook). D'aquesta manera en un sol pas obtenim tots els valors diferents de rangs d'edat de l'origen de dades.
2. Les dades obtingudes al pas anterior s'injecten a la taula de dimensió *Age* del DW (es vinculen els camps i s'injecten les dades).
3. En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Age,* ja té dades que poden ser utilitzades.

4.3.7 Generació de la dimensió *Gender*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Gender* es mostra a continuació:

Figura 24: Transformació de generació de la dimensió *Gender*

Els passos de transformació per a la obtenció de les dades per a la dimensió *Gender*, es detallen a continuació:

- S'obtenen tots els generes diferents de les taules de dades amb una consulta SQL. La consulta uneix els resultats de les taules de dades (Youtube, Instagram, Twitter i Facebook). D'aquesta manera en un sol pas obtenim tots els valors diferents de sexes de l'origen de dades.
- Les dades obtingudes al pas anterior s'injecten a la taula de dimensió *Gender* del DW (es vinculen els camps i s'injecten les dades).
- En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Gender,* ja té dades que poden ser utilitzades.

4.3.8 Generació de la dimensió *Likes*

La implementació de la transformació per a la generació de les dades de la taula de dimensió *Likes* es mostra a continuació:

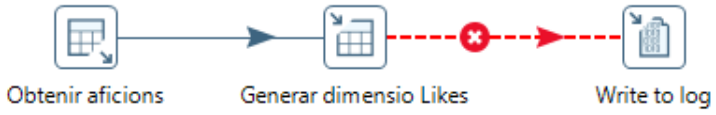


Figura 25: Transformació de generació de la dimensió *Likes*

Els passos de transformació per a la obtenció de les dades per a la dimensió *Likes*, es detallen a continuació:

1. S'obtenen totes les aficions diferents de les taules de dades amb una consulta SQL. La consulta uneix els resultats de les taules de dades (Youtube, Instagram, Twitter i Facebook). D'aquesta manera en un sol pas obtenim tots els valors diferents d'aficions de l'origen de dades.
2. Les dades obtingudes al pas anterior s'injecten a la taula de dimensió *Likes* del DW (es vinculen els camps i s'injecten les dades).
3. En cas d'error es redirigeix la sortida de la transformació al fitxer de log.

Un cop s'ha finalitzat la transformació la taula de dimensió *Likes,* ja té dades que poden ser utilitzades.

4.3.9 Generació de la taula de Fets *CRT*

La implementació d'aquesta transformació està dividida en dues parts: una transformació pare i una transformació filla. La transformació pare, a partir de les diferents xarxes socials (dimensió *Network*) fa crides a la transformació filla perquè executi els seus passos per la xarxa social que el pare indiqui. Com que hi ha quatre xarxes socials, el pare cridarà quatre vegades la transformació filla (una vegada per a cada xarxa social). La transformació filla realitzarà els passos necessaris per obtenir la informació relativa a una xarxa social en el format correcte per poder ser injectada al DW. A continuació es mostra la implementació de la transformació pare:

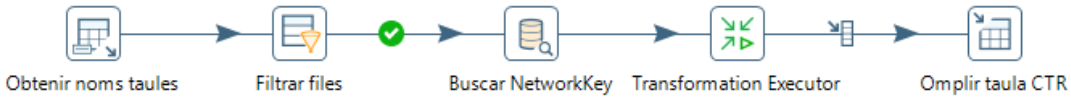


Figura 26: Transformació pare de generació de la taula de fets

Els passos d'aquesta transformació són els següents:

1. Obtenim els noms de les taules del nostre origen de dades (és una BD MySQL).
2. Filtrem la informació i ens quedem amb els noms de taula que es corresponen a una xarxa social (Youtube, Instagram, Twitter i Facebook).

3. Com que el DW ja té la dimensió Network creada, per a cada xarxa social cerquem l'identificador generat (*NetworkKey*).
4. Passem l'identificador de xarxa social i el nom a la transformació filla mitjançant un pas "Transformation Executor". Aquest pas s'executa tantes vedades com registres hi hagi a l'stream de dades que rep. En el nostre cas l'stream tindrà quatre registres (un per a cada xarxa social).
5. Finalment, amb les dades que rebrà de la transformació filla, s'omplirà la taula de fets (es vinculen els camps i s'injecten les dades).

A continuació es mostra la implementació de la transformació filla (que obté les dades per inserir a la taula de fets):

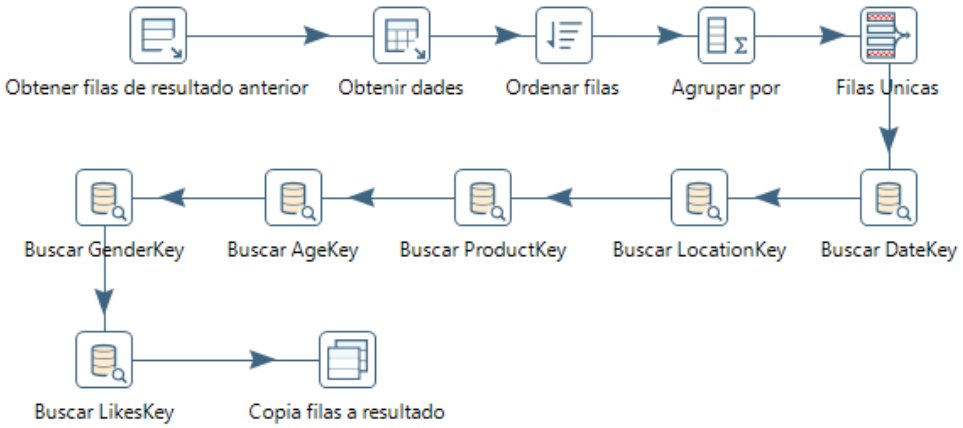


Figura 27: Transformació filla de generació de la taula de fets *CTR*

Els passos que es realitzen són els següents:

1. Obtenim la informació que prové d'un resultat anterior, en el nostre cas prové de la transformació pare. Els resultats que s'obtenen són el codi i nom de xarxa social que es tractarà.
2. Amb la informació del pas anterior s'executa una consulta SQL contra l'origen de dades per obtenir tota la informació relacionada amb la xarxa social a tractar.
3. S'ordenen les files per data, producte, localització, edat, sexe i aficions.
4. S'agrupen les dades per data, producte, localització, edat, sexe i aficions. Es calculen els camps suma d'impressions i suma de clics en funció de l'agrupació. Aquesta agrupació, tot i que a priori no és necessària, ens assegura valors únics de registre.
5. Es consideren les files úniques de l'agrupació anterior.
6. Es busca la clau que es correspon a la data dels registres del DataSet (*DateKey*).
7. Es busca la clau que es correspon a la localització dels registres del DataSet (*LocationKey*).
8. Es busca la clau que es correspon al producte dels registres del DataSet (*ProductKey*).
9. Es busca la clau que es correspon al rang d'edats dels registres del DataSet (*AgeKey*).
10. Es busca la clau que es correspon al gènere dels registres del DataSet (*GenderKey*).

11. Es busca la clau que es correspon a les aficions dels registres del DataSet (*LikesKey*).
12. Tenim tota la informació amb el format desitjat (codis enlloc de literals, impressions i clics agrupats) llesta per ser injectada al DW. La deixem disponible com a resultat de la transformació.

El resultat de la transformació filla es rebut pel pare, que injecta de forma efectiva les dades a la taula de fets del DW.

Tot i que les transformacions implementades no tenen un alt grau de complexitat, ha sigut necessari disposar d'una font d'informació i consulta fiable. Per a la realització d'aquesta part del TFM s'ha utilitzat com a bibliografia bàsica i font de referència la Wiki de Pentaho [26] així com [27].

4.4 Comprovacions de qualitat de les dades del DW

Un cop les transformacions s'han executat sense errors, és necessari realitzar una comprovació general de que les dades que s'han injectat al DW es corresponen amb la informació que hi havia a l'origen de dades.

La validació de les taules de dimensió es pot fer de forma visual explorant les dades que s'han injectat. Si a les taules del DW hi ha poques dades, la comparació es pot fer de manera senzilla sense necessitat de recórrer a la realització d'altres processos de validació. Per a les dimensions *Product*, *Location*, *Age*, *Likes* i *Gender*, es determina que la correspondència entre els orígens de dades i les taules de destí són correctes un cop s'han aplicat les transformacions.

La validació de les dades de la taula de dimensió *Date*, no és tan directa com les altres dimensions. La taula de dimensió té un total de 365 registres que, efectivament, es corresponen als 365 dies de l'any 2017. Una validació addicional que s'ha de realitzar és comprovar aquells dies en els que hi ha un canvi de mes, trimestre o semestre i confirmar que les dades que contenen els registres són correctes. A continuació es mostren aquestes validacions:

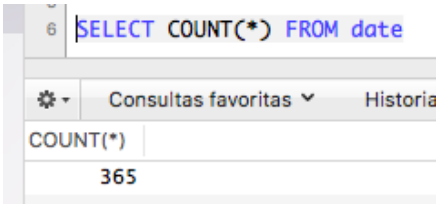


Figura 28: Total de registres a la dimensió *Date*

2017-01-29	29	Sunday	January	Q1	S1	2017
2017-01-30	30	Monday	January	Q1	S1	2017
2017-01-31	31	Tuesday	January	Q1	S1	2017
2017-02-01	1	Wednesday	February	Q1	S1	2017
2017-02-02	2	Thursday	February	Q1	S1	2017

Figura 29: Validació de les dades en un canvi de mes (de Gener a Febrer)

2017-03-30	30	Thursday	March	Q1	S1	2017
2017-03-31	31	Friday	March	Q1	S1	2017
2017-04-01	1	Saturday	April	Q2	S1	2017
2017-04-02	2	Sunday	April	Q2	S1	2017

Figura 30: Validació de les dades en un canvi de trimestre (de Març a Abril)

2017-06-29	29	Thursday	June	Q2	S1	2017
2017-06-30	30	Friday	June	Q2	S1	2017
2017-07-01	1	Saturday	July	Q3	S2	2017
2017-07-02	2	Sunday	July	Q3	S2	2017

Figura 31: Validació de les dades en un canvi de semestre (de Juny a Juliol)

La taula que conté més dades i sobre la que s'han fet operacions d'agrupació durant la transformació és la taula de fets. Per poder validar que les dades que s'han injectat es corresponen de forma fidel amb les dades originals ens ajudarem de consultes SQL en una i altre BD per tal de confirmar que els resultats són els mateixos.

En primer lloc comprovarem que el total de files úniques (agrupades pels valors clau de la taula de fets) es corresponen entre l'origen de dades i el DW. Primer consultarem quants registres únics hi ha la BD d'origen.

Xarxa social	Registres	Total prints	Total hits
facebook	373760	26437475	2600754
instagram	373760	26444417	2662052
twitter	373760	26449815	2599397
youtube	373760	26380191	2664664

Figura 32: Dades totals per xarxa social a l'origen de dades

A continuació consultarem quants registres hi ha a la taula de fets per cada xarxa social (recordem que a la taula de fets les dades ja estan agrupades):

Xarxa social	Registres	Total prints	Total hits
facebook	373760	26437475	2600754
instagram	373760	26444417	2662052
twitter	373760	26449815	2599397
youtube	373760	26380191	2664664

Figura 33: Dades totals per xarxa social a la taula de fets del DW

Com es pot observar el resultat és el mateix, i per tant en una primera validació, podem dir que el total de dades (en conjunt) és el mateix tant en l'origen de dades com en el DW.

S'han realitzat comprovacions addicionals per validar que les dades del DW són correctes; per exemple s'han analitzat de forma global els dos semestres i les dades de cada més de l'any 2017. En aquest procés d'anàlisi s'ha pogut validar que les dades globals s'han traspassat en el mateix nombre de l'origen de dades al DW.

A continuació es mostra (a mode d'exemple) una de les validacions realitzades per comprovar que, a nivell de de codi postal, el total de dades migrades i el valor d'impressions i clics coincideix entre l'origen de dades i el DW:

Codi postal	Registres	Total prints	Total hits
08005	93440	7600411	860336
08029	93440	7610939	861570
11011	93440	6099472	489193
15010	93440	6103785	519965
28014	93440	7629428	864734
28019	93440	7618697	867026
29009	93440	6480185	592644
33209	93440	5705302	488470
41011	93440	6858341	671841
41092	93440	6867309	672348
43006	93440	6471654	666068
45005	93440	4957407	380472
46025	93440	7041608	736292
47011	93440	5324761	392621
48003	93440	6686079	734267
48008	93440	6656520	729020

Figura 34: Dades totals per codi postal a l'origen de dades.

Codi postal	Registres	Total prints	Total hits
08005	93440	7600411	860336
08029	93440	7610939	861570
11011	93440	6099472	489193
15010	93440	6103785	519965
28014	93440	7629428	864734
28019	93440	7618697	867026
29009	93440	6480185	592644
33209	93440	5705302	488470
41011	93440	6858341	671841
41092	93440	6867309	672348
43006	93440	6471654	666068
45005	93440	4957407	380472
46025	93440	7041608	736292
47011	93440	5324761	392621
48003	93440	6686079	734267
48008	93440	6656520	729020

Figura 35: Dades totals per codi postal a la taula de fets del DW

5. Disseny i implementació de cubs OLAP per a l'explotació de la informació del DW.

5.1 Definició i tipologia de cubs OLAP

Entenem per cubs OLAP (OnLine Analytical Processing), a aquelles estructures de dades organitzades de tal manera que poden emmagatzemar una representació complexa de dades orientades a facilitar capacitat d'anàlisi de forma ràpida i eficient. Els cubs OLAP són multidimensionals, permeten fer operacions ràpidament sobre grans conjunts de dades i, a més, proporcionen capacitat als usuaris per poder accedir al detall de les dades o, al contrari, obtenir una visió més general. Permeten accions de filtratge i ordenació avançats de manera que la visualització de les dades s'adapti a les necessitats dels usuaris finals.

Històricament les dades emmagatzemades en sistemes de BD tradicionals, oferien característiques més que suficients per ajudar a les empreses a analitzar-les i prendre decisions. Amb el pas del temps les necessitat s'han tornat més i més exigents a mesura que la quantitat de dades recopilades amb informació valuosa creixia. Davant aquesta situació els sistemes tradicionals de BD relacionals tenen mancances evidents ja que els hi costa gestionar grans quantitats de dades per realitzar anàlisis complexos de forma eficient. Aquestes limitacions són superades gràcies a la utilització de cubs OLAP.

Un sistema OLAP (que de forma genèrica s'anomena cub OLAP) està format per diversos components: un origen de dades, un servidor OLAP i un client. L'origen de dades és precisament el contenidor de les dades que es volen analitzar, que en aquest cas es tracta del DW de les dades de publicitat. El servidor OLAP, o back-end, és on es realitza la feina de consulta i processament de la informació, és el nucli del sistema. Finalment el client, és l'eina o eines que s'usen per tal de visualitzar la informació processada [28].

Existeixen diferents tipus de sistemes OLAP ben diferenciats entre ells. El primer tipus és que rep el nom d'MOLAP (Multi-dimensional OnLine Analytical Processing). En aquests tipus de cubs el motor de BD sobre el que treballa el sistema és un motor de base de dades multidimensional. Els cubs MOLAP tenen avantatges, per exemple el seu rendiment és molt bo, però per contra els processos de càrrega de dades a les BD multidimensionals són complexos induint (en ocasions) a possibles duplicitats de dades.

En el sentit oposat als cubs MOLAP trobem els cubs ROLAP (Relational OnLine Analytical Processing) Aquests cubs fan servir com a motor de BD sistemes relacionals on els esquemes de BD utilitzats són de tipus estrella o floc de neu. Comparant-los amb els cubs MOLAP, aquests són més lents però molt útils a l'hora de processar grans volums de dades.

Finalment trobem solucions híbrides que tracten de combinar el millor de cada sistema per potenciar les virtuts i minimitzar els inconvenients. Aquests tipus de cubs reben el nom d'HOLAP (Hybrid OnLine Analytical Processing). Aquests cubs emmagatzemen una part de les dades en format MOLAP per tenir alta velocitat de resposta i la resta de dades en format ROLAP per facilitat d'esquema i emmagatzemament.

Un cop exposats els diferents tipus de cubs OLAP, comentar que el tipus de cub OLAP emprat en aquest TFM, és un cub ROLAP, ja que el motor de BD es relacional i l'esquema del DW dissenyat segueix el model en estrella.

5.2 Definició del cub OLAP

Per tal de poder analitzar de forma còmode i senzilla la informació emmagatzemada al DW, es dissenyarà un cub OLAP (ROLAP) que ens permeti visualitzar la informació rellevant,

ajustar el nivell de detall de les dades que es mostren i donar resposta a les preguntes analítiques formulades.

La plataforma Pentaho disposa d'una eina per a la creació i gestió de cubs OLAP que s'anomena Schema Workbench. El seu funcionament és senzill i permet la creació de tants cubs com siguin necessaris, així com la seva publicació a la plataforma BI Pentaho, de manera que els cubs estaran disponibles per als clients que necessitin analitzar les dades exposades pels cubs. En el cas d'estudi, es defineix un únic cub que serà el que s'utilitzarà per a tot l'anàlisi de dades. El cub creat es facilita al fitxer "DefinicioCubOLAP.xml".

No és l'objectiu d'aquest treball explicar quin és el procediment de creació del cub OLAP. Per aquells lectors interessats en la generació i definició de cubs sobre la plataforma BI Pentaho, es pot consultar la referència bibliogràfica [29] (juntament amb multitud de recursos disponibles a Youtube). A continuació es mostra la definició del cub OLAP que es farà servir per a l'anàlisi de les dades i les seves característiques més importants.

En primer lloc, i com que només definirem un cub per a la realització de tot l'anàlisi és necessari definir les dimensions d'anàlisi que tindrem en compte. Aquesta definició no forma part del cub, però és un pas previ totalment necessari per poder crear el cub posteriorment. Com es pot observar es creen totes les dimensions susceptibles de ser utilitzades en l'anàlisi posterior:

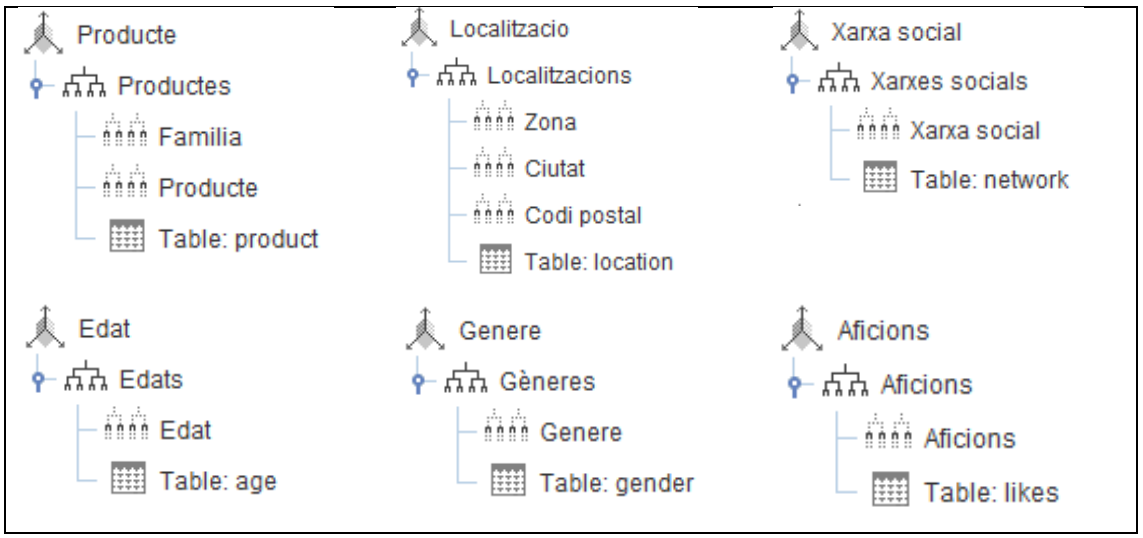


Figura 36: Definició de dimensions a Pentaho Schema Workbench

Com es pot observar en la figura anterior, la definició de les dimensions conté la taula de dimensions del DW que serveix d'origen de dades a cadascuna de les dimensions i una jerarquia d'elements. En tots els casos anteriors trobem una única jerarquia d'elements. Dins de cada jerarquia es defineixen els nivells. Es pot observar que les dimensions *Producte* i *Localització* tenen dos nivells (dins la mateixa jerarquia), això significa que en el moment de fer l'anàlisi de dades sobre aquestes dimensions, es podrà arribar a un nivell de detall diferent. Per exemple, a *Productes* podrem analitzar les dades a nivell de *Família* o a nivell de *Producte*. Notar que els nivells s'estructuren de més genèric a més detallat (*Zona-Ciutat-Codi postal*).

Finalment la darrera dimensió és la dimensió de Calendari, que ens permetrà fer una anàlisi de tendències al llarg del temps. Per la definició d'aquesta dimensió definim varies jerarquies que es mostren a continuació:

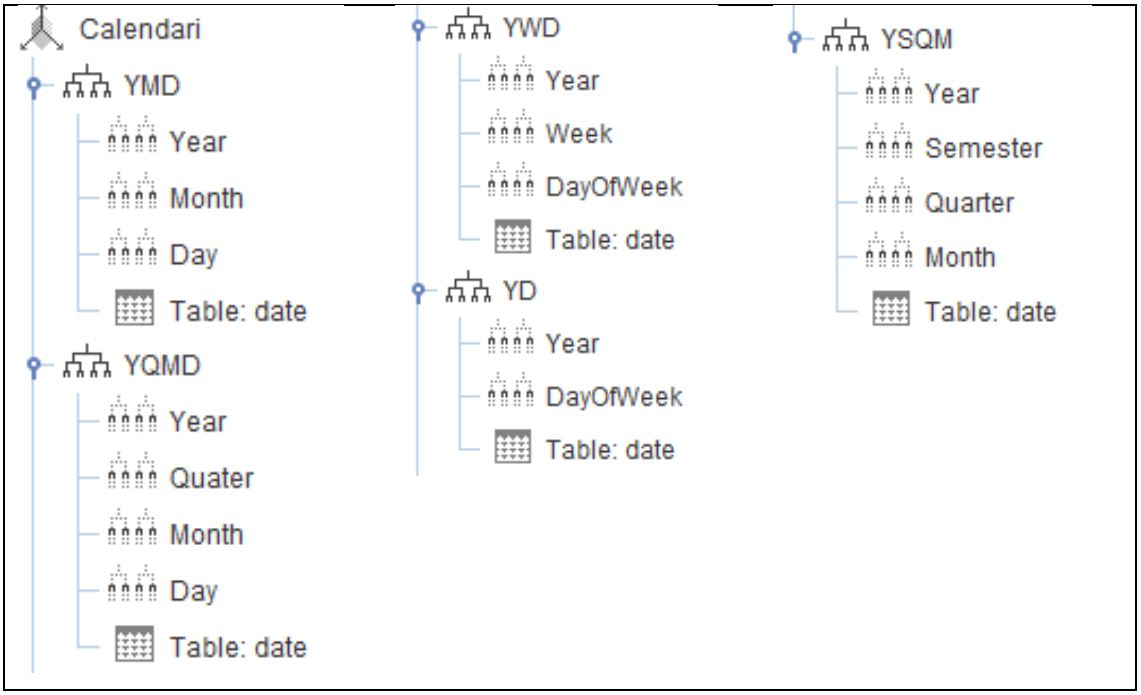


Figura 37: Definició de la dimensió Calendari

Un cop definides les dimensions, es crea el cub OLAP. En el cas que ens ocupa, el cub contindrà totes les dimensions, actuarà sobre la taula CTR del DW, contindrà dues mesures i una mesura calculada. Les mesures (en un cub OLAP) són aquelles dades sobre les que farem l'anàlisi dimensional. El nostre cub tindrà les mesures d'impressions i clics, i a més la mesura calculada CTR (que és la que realment ens interessa per a l'anàlisi). El CTR és una mesura calculada perquè no es correspon a cap informació de la taula de fets i el seu valor s'obté a partir de càlculs sobre les mesures d'impressions i de clics. A continuació es mostra la definició del cub OLAP:

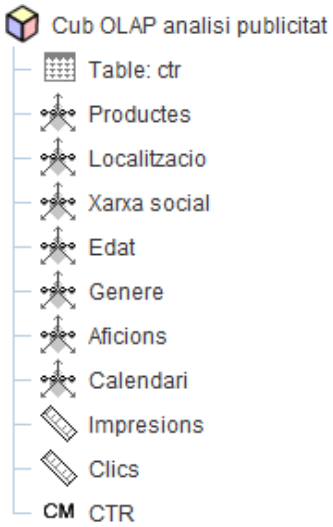


Figura 38: Definició de cub OLAP per a l'anàlisi de la publicitat

6. Anàlisi de dades del DW a partir de cubs OLAP

6.1 Elements per a la realització de l'anàlisi de les dades de publicitat

Les plataformes BI, ofereixen als usuaris potents eines per a la visualització i tractament de les dades multidimensionals, i Pentaho no és una excepció. Pentaho integra en la seva instal·lació per defecte, JPivot que permet realitzar les operacions típiques sobre cubs OLAP (Slice, Dice, Drill Down i Roll Up), permetent l'anàlisi de la informació continguda a la representació del cub. Malgrat tot, JPivot és una eina antiga amb una interfície d'usuari que actualment no resulta atractiva per al seu ús i funcionalitats limitades. Pentaho disposa d'un ric Marketplace en el que es poden trobar moltes aplicacions i plugins per a la plataforma de forma gratuïta.

Per a l'anàlisi i visualització de dades s'ha triat una d'aquestes aplicacions: Saiku Business Analytics, que proporciona una versió Community Edition del seu producte i que millora les prestacions de JPivot en quant a facilitat d'ús, generació de gràfics que proporcionen una visualització còmode de les dades, possibilitat de generar Dashboards de control, etc.

Un cop descarregada l'aplicació i afegida a Pentaho, el seu funcionament és molt senzill i intuïtiu. Es selecciona el cub a utilitzar i automàticament apareixen les mesures i dimensions que, mitjançant operacions de Drag&Drop, s'afegeixen a la configuració de la visualització de dades amb la que es vol treballar.

A continuació es presenten les conclusions que es poden obtenir de l'anàlisi de les dades de publicitat del DW en base a les preguntes formulades.

6.2 Anàlisi de dades i resposta a les preguntes analítiques proposades.

L'indicador bàsic utilitzat per analitzar les dades de publicitat és el CTR. Aquest indicador ens dóna una idea de l'efectivitat dels anuncis, al mesurar el percentatge de clics que es realitzen en relació a les vegades que es mostren els anuncis per pantalla. És un indicador que no és sensible a variacions brusques del número d'impressions, és a dir que no guarda cap tipus de proporcionalitat amb el número de vegades que un anunci es mostra per pantalla. Per aquest motiu és un indicador molt vàlid a partir del qual es poden extreure conclusions.

Cal tenir en compte que un anàlisi més efectiu es podria haver dut a terme si, a més de CTR, les dades del problema ens haguessin ofert el percentatge de conversions. En màrqueting online, les campanyes publicitàries s'enfoquen a l'obtenció de conversions. Una conversió pot ser una venda d'un eCommerce, un registre d'usuari a un servei, obtenir el e-mail dels usuaris per campanyes de mailing o, fins i tot que els usuaris accedeixin a un contingut específic sense generar un rebot (es considera un rebot aquell usuari que després de visitar el contingut publicitat no realitza cap interacció i abandona el lloc web). Les possibles dades de conversions haguessin aportat un punt de vista més acorat del que ha passat amb els anuncis i permetrien avaluar amb més profunditat, si aquests anuncis són efectius (compleixen amb el seu objectiu) o no. Les dades que proporcionades, serviran per poder determinar si les persones que veuen els anuncis reaccionen positivament a ells (sigui per necessitat, curiositat, etc.) i fan clic.

Feta aquesta observació passem a l'anàlisi de les dades en funció de les preguntes analítiques proposades.

6.2.1 Què regions o ciutats tenen millors indicadors d'efectivitat? Hi ha alguna relació amb el producte o família de productes?

Per donar resposta a la primera qüestió, mostrarem el valor del CTR en funció de la localització des de la que es va clicar als anuncis (files). Anirem afegint detall a l'anàlisi passant de zona a ciutat i a codi postal. El resultat de les taules de visualització de dades es mostren a continuació:

Zona	CTR
Centre	9,81%
Coast	10,88%
North	9,83%
South	9,22%

Zona	Ciutat	CTR
Centre	Madrid	11,36%
	Toledo	7,67%
	Valladolid	7,37%
Coast	Barcelona	11,32%
	Tarragona	10,29%
	Valencia	10,46%
North	Bilbao	10,97%
	Gijon	8,56%
	La Coruña	8,52%
South	Cadiz	8,02%
	Malaga	9,15%
	Sevilla	9,79%

Zona	Ciutat	Codi postal	CTR
Centre	Madrid	28014	11,33%
		28019	11,38%
	Toledo	45005	7,67%
	Valladolid	47011	7,37%
Coast	Barcelona	08005	11,32%
		08029	11,32%
	Tarragona	43006	10,29%
	Valencia	46025	10,46%
North	Bilbao	48003	10,98%
		48008	10,95%
	Gijon	33209	8,56%
	La Coruña	15010	8,52%
South	Cadiz	11011	8,02%
	Malaga	29009	9,15%
	Sevilla	41011	9,80%
		41092	9,79%

Figura 39: Rendiment en funció de la localització (Zona / Zona-Ciutat / Zona-Ciutat-CP)

Analitzant per zona geogràfica global, es veu que el millor rendiment s'obté a *Coast* amb un increment del CTR d'un 1,07% respecte el següent (*North*). Els valors són tots molt semblants. Si analitzem les dades afegint més detall (ciutats) veiem que la tendència (com es podria esperar) es manté però obtenim informació addicional força interessant.

Les ciutats de la zona *Coast* tenen un comportament similar entre elles, destacant per sobre de la resta *Barcelona*. A la zona *Centre* veiem que *Madrid* és la ciutat que millor respon, però hi ha una diferència molt marcada entre *Madrid* i les altres ciutats de la zona (3,69%) A la zona *North* passa el mateix que el que es descriu per *Centre*. Finalment a *South* tenim un comportament similar a *Coast*. Quan s'analitzen les dades a nivell de codi postal, s'observa que aquest nivell de detall no aporta cap singularitat o dada remarcable.

Es pot concloure que, efectivament hi ha zones geogràfiques on els indicadors són clarament millors. Es rellevant el cas de les zones *Centre* i *North* on *Madrid* i *Bilbao* responen força millor que les altres localitzacions de la seva zona. Si s'hagués de maximitzar la inversió potser caldria evitar mostrar els anuncis a les altres ciutats, mostrant-los només a *Madrid* i *Bilbao*. També podem concloure que les ciutats amb més densitat de població són aquelles on millor responen els usuaris als anuncis (*Madrid*, *Barcelona*, *Bilbao* i *Sevilla*).

A continuació es mostren les dades de les taules en format gràfic:

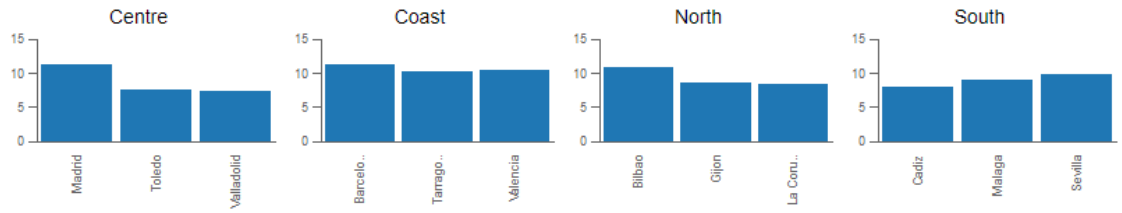


Figura 40: Gràfic del CTR en funció de la localització

Per donar resposta a la segona pregunta, s'inclou la dimensió de productes (columnes) a la taula d'anàlisi de dades. En una primera aproximació s'inclou només la família de productes per analitzar la tendència. El resultat és el següent (amb i sense codi postal a l'anàlisi):

	Familia	Accessory	Culture	Electronics	Sports	Wear
Zona	Ciutat	CTR	CTR	CTR	CTR	CTR
Centre	Madrid	11,34%	11,36%	11,36%	11,39%	11,35%
	Toledo	7,68%	7,69%	7,65%	7,71%	7,65%
	Valladolid	7,39%	7,33%	7,40%	7,36%	7,39%
Coast	Barcelona	11,40%	11,32%	11,25%	11,26%	11,31%
	Tarragona	10,31%	10,30%	10,26%	10,33%	10,25%
	Valencia	10,38%	10,46%	10,55%	10,48%	10,47%
North	Bilbao	10,95%	10,95%	10,93%	10,96%	11,03%
	Gijon	8,55%	8,56%	8,58%	8,45%	8,62%
	La Coruña	8,48%	8,55%	8,59%	8,52%	8,50%
South	Cadiz	8,04%	8,00%	8,15%	7,96%	7,99%
	Malaga	9,18%	9,17%	9,13%	9,12%	9,10%
	Sevilla	9,85%	9,77%	9,85%	9,81%	9,72%

	Familia	Accessory	Culture	Electronics	Sports	Wear
Zona	Ciutat	Codi postal	CTR	CTR	CTR	CTR	CTR
Centre	Madrid	28014	11,28%	11,35%	11,38%	11,32%	11,35%
		28019	11,39%	11,38%	11,35%	11,46%	11,35%
	Toledo	45005	7,68%	7,69%	7,65%	7,71%	7,65%
	Valladolid	47011	7,39%	7,33%	7,40%	7,36%	7,39%
Coast	Barcelona	08005	11,45%	11,34%	11,31%	11,19%	11,23%
		08029	11,34%	11,30%	11,20%	11,33%	11,38%
	Tarragona	43006	10,31%	10,30%	10,26%	10,33%	10,25%
	Valencia	46025	10,38%	10,46%	10,55%	10,48%	10,47%
North	Bilbao	48003	10,98%	10,94%	10,93%	10,96%	11,07%
		48008	10,92%	10,96%	10,93%	10,96%	10,99%
	Gijon	33209	8,55%	8,56%	8,58%	8,45%	8,62%
	La Coruña	15010	8,48%	8,55%	8,59%	8,52%	8,50%
South	Cadiz	11011	8,04%	8,00%	8,15%	7,96%	7,99%
	Malaga	29009	9,18%	9,17%	9,13%	9,12%	9,10%
	Sevilla	41011	9,86%	9,77%	9,84%	9,79%	9,75%
		41092	9,85%	9,77%	9,86%	9,84%	9,70%

Figura 41: Taules d'anàlisi de rendiment per localització i família

Com es pot observar no hi ha cap canvi respecte el rendiment sense tenir en compte la família. Els valors es mantenen estables tant per zona, per ciutat i per codi postal.

Si afegim més detall incloent els productes de les famílies (més detall a nivell de columnes) obtenim la següent taula d'anàlisi de dades (sense arribar al detall de codi postal per a la dimensió de localització):

	Familia	Accessory		Culture		Electronics	Sports	Wear	
	Producte	Scarf	Watch	Theater	Trip	Mobile Phone	Sneakers	Dress	Sheatshirt
Zona	Ciutat	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
Centre	Madrid	11,33%	11,34%	11,33%	11,39%	11,36%	11,39%	11,43%	11,27%
	Toledo	7,62%	7,74%	7,72%	7,66%	7,65%	7,71%	7,55%	7,76%
	Valladolid	7,37%	7,41%	7,30%	7,37%	7,40%	7,36%	7,32%	7,46%
Coast	Barcelona	11,36%	11,43%	11,32%	11,32%	11,25%	11,26%	11,36%	11,26%
	Tarragona	10,25%	10,38%	10,32%	10,29%	10,26%	10,33%	10,24%	10,26%
	Valencia	10,40%	10,36%	10,44%	10,48%	10,55%	10,48%	10,48%	10,46%
North	Bilbao	10,93%	10,97%	10,97%	10,93%	10,93%	10,96%	11,12%	10,93%
	Gijon	8,55%	8,55%	8,62%	8,51%	8,58%	8,45%	8,64%	8,60%
	La Coruña	8,45%	8,50%	8,54%	8,55%	8,59%	8,52%	8,56%	8,44%
South	Cadiz	8,10%	7,98%	7,92%	8,07%	8,15%	7,96%	7,89%	8,09%
	Malaga	9,17%	9,18%	9,16%	9,19%	9,13%	9,12%	9,10%	9,10%
	Sevilla	9,84%	9,86%	9,74%	9,80%	9,85%	9,81%	9,72%	9,72%

Figura 42: Taula d'anàlisi de rendiment en funció de la localització i productes

Les conclusions són les mateixes que les obtingudes anteriorment. A nivell de localització no es pot establir cap relació amb la família ni amb el producte anunciat. Podem concloure que els anuncis funcionen (aproximadament) amb la mateixa efectivitat per zones geogràfiques independentment de la família o producte que s'estigui anunciant. Quan a una zona la publicitat té un CTR alt (*Madrid* per exemple) el continua tenint alt, amb independència del producte o família de productes anunciat (les variacions són molt petites).

6.2.2 Existeix una relació entre la millora dels indicadors d'efectivitat amb algun segment de la població objectiu?

Sobre la consulta de visualització de dades s'aniran realitzant certes comprovacions de rendiment de publicitat (CTR) en funció de les característiques del públic objectiu (dimensions sexe edat i aficions). Inicialment s'analitza el rendiment per grups d'edat i sexe; els resultats obtinguts revelen que la resposta envers els anuncis no es veu afectada pel sexe del públic. El que si s'observa és que clarament el grup d'usuaris que té una edat compresa entre els 61-99 anys no respon igual de bé que la resta de les franges d'edat (quasi un 3% menys).

Genere	Female	Male
Edat	CTR	CTR
18-30	10,09%	10,07%
31-45	10,55%	10,54%
46-60	10,56%	10,57%
61-99	8,61%	8,66%

Figura 43: Taula d'anàlisi de rendiment en funció de l'edat i el sexe

Si afegim a l'anàlisi la tercera característica important en la definició del públic objectiu (les seves aficions) veiem com que tampoc hi ha cap diferència notable entre homes i dones en relació al CTR:

	Genere	Female	Male
Edat	Aficions	CTR	CTR
18-30	Business	7,68%	7,71%
	Cars	10,82%	10,75%
	Fashion	10,87%	10,83%
	Garden	6,96%	6,97%
	People	12,97%	12,94%
	Sports	11,23%	11,30%
	Technology	8,89%	8,90%
	Travel	11,24%	11,24%

31-45	Business	8,16%	8,14%
	Cars	11,28%	11,26%
	Fashion	11,38%	11,24%
	Garden	7,47%	7,42%
	People	13,45%	13,49%
	Sports	11,70%	11,73%
	Technology	9,37%	9,34%
	Travel	11,72%	11,72%

46-60	Business	8,20%	8,19%
	Cars	11,28%	11,25%
	Fashion	11,23%	11,41%
	Garden	7,45%	7,43%
	People	13,39%	13,37%
	Sports	11,68%	11,75%
	Technology	9,43%	9,42%
	Travel	11,80%	11,74%

61-99	Business	6,29%	6,25%
	Cars	9,26%	9,44%
	Fashion	9,28%	9,35%
	Garden	5,51%	5,55%
	People	11,47%	11,61%
	Sports	9,70%	9,84%
	Technology	7,46%	7,46%
	Travel	9,89%	9,78%

Figura 44: Taula d'anàlisi de rendiment en funció de l'edat, el sexe i les aficions

Una vegada confirmat que el sexe no és un factor clau en aquest anàlisi, el podem eliminar per centrar-nos en veure les relacions que es poden establir entre grups d'edat i aficions:

Aficions	Business	Cars	Fashion	Garden	People	Sports	Technology	Travel
Edat	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
18-30	7,69%	10,78%	10,85%	6,97%	12,96%	11,26%	8,89%	11,24%
31-45	8,15%	11,27%	11,31%	7,44%	13,47%	11,72%	9,36%	11,72%
46-60	8,19%	11,26%	11,32%	7,44%	13,38%	11,72%	9,42%	11,77%
61-99	6,27%	9,35%	9,32%	5,53%	11,54%	9,77%	7,46%	9,83%

Figura 45: Taula d'anàlisi de rendiment en funció de l'edat i aficions

Com es pot observar en la taula de dades (Figura 45), es poden establir les relacions següents:

1. Les persones que millor responen als anuncis (sense tenir en compte la temàtica de l'anunci) són aquelles que tenen per aficions: *People*, *Sports* i *Travel*. El públic que pitjor reacciona als anuncis és aquell amb afició a *Garden*.
2. Les franges d'edat que millor responen als anuncis son: 31-45 i 46-60. Clarament la pitjor franja és 61-99, on obtenim les pitjors dades de rendiment per a qualssevol afició. En aquestes dues franges d'edat (31-45 i 46-60) el rendiment és pràcticament el mateix.
3. El rendiment és bastant heterogeni, és a dir que un CTR alt en una franja d'edat i afició normalment comporta que el CTR vinculat a l'afició analitzada i per qualssevol franja d'edat també serà alt, en relació a les altres aficions.
4. Notar que destaca positivament el CTR per a la combinació 61-99 amb afició *People* amb un 11,54% inferior a les altres franges d'edat però el millor CTR per afició en aquesta franja d'edat.

Extreure una conclusió sobre com maximitzar el rendiment en base a aquestes dades, és complicat. El que es veu clar és que seria positiu evitar mostrar anuncis a persones de més de 60 anys, ja que el rendiment està força per sota del de la resta de franges. Aventurar-se a prendre més decisions sense incloure altres dimensions, no sembla el més correcte, ja que tot i tenir un CTR baix, pot ser que una combinació d'edat i afició cobreixi un nínxol de mercat important per l'empresa i per tant s'hauria d'avaluar en base a més paràmetres. A continuació es mostren els resultats de forma gràfica:

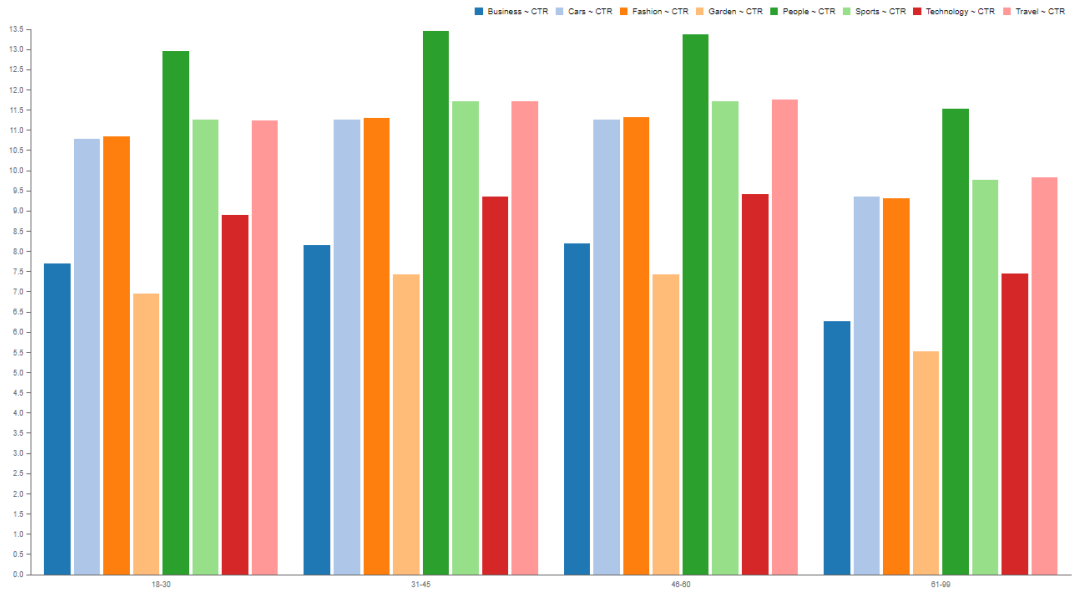


Figura 46: Representació gràfica del rendiment en funció d'edat i aficions

6.2.3 Hi ha alguna plataforma, on sota les mateixes condicions, s'obtinguin millors taxes de visualització?

Iniciarem el procés d'anàlisi centrant-nos en el rendiment dels diferents productes en funció de la xarxa social sobre la que es publica l'anunci. Les dades de l'anàlisi es mostren a continuació:

	Xarxa social	facebook	instagram	twitter	youtube
Familia	Producte	CTR	CTR	CTR	CTR
Accessory	Scarf	9,82%	10,14%	9,78%	10,09%
	Watch	9,89%	10,09%	9,84%	10,12%
Culture	Theater	9,82%	10,04%	9,87%	10,05%
	Trip	9,88%	10,11%	9,81%	10,06%
Electronics	Mobile Phone	9,85%	10,07%	9,83%	10,13%
Sports	Sneakers	9,78%	10,05%	9,81%	10,15%
Wear	Dress	9,84%	10,04%	9,88%	10,12%
	Sheatshirt	9,81%	9,99%	9,81%	10,09%

Figura 47: Taula d'anàlisi de productes en funció de la xarxa social

Les dades ens mostren que, de forma genèrica, tots els productes tenen un rendiment homogeni per xarxa social. Entre xarxes socials, si que s'observen canvis pel mateix producte, per exemple *Theater* respon millor a Instagram que a Twitter. En general, a Instagram i Youtube els anuncis responen una mica millor que a Facebook i Twitter.

Ara ens preguntem si el sexe pot ser un factor a tenir en compte; i l'afegim a l'anàlisi. En base a les dades mostrades a la taula que es presenta continuació, podem concloure que el sexe no és un factor que faci variar el rendiment dels anuncis.

	Xarxa social	facebook		instagram		twitter		youtube	
	Genere	Female	Male	Female	Male	Female	Male	Female	Male
Familia	Producte	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
Accessory	Scarf	9,83%	9,82%	10,16%	10,11%	9,75%	9,80%	10,10%	10,07%
	Watch	9,97%	9,82%	10,06%	10,13%	9,89%	9,78%	10,11%	10,12%
Culture	Theater	9,83%	9,81%	10,00%	10,09%	9,90%	9,84%	10,02%	10,09%
	Trip	9,88%	9,88%	10,06%	10,15%	9,79%	9,82%	10,06%	10,06%
Electronics	Mobile Phone	9,82%	9,87%	10,08%	10,05%	9,86%	9,80%	10,10%	10,15%
Sports	Sneakers	9,81%	9,75%	9,97%	10,13%	9,73%	9,89%	10,16%	10,14%
Wear	Dress	9,85%	9,84%	10,02%	10,07%	9,90%	9,86%	10,17%	10,07%
	Sheatshirt	9,76%	9,87%	9,92%	10,06%	9,86%	9,76%	10,08%	10,10%

Figura 48: Taula d'anàlisi en funció de productes, xarxa social i edat

Si afegim a l'anàlisi la dimensió d'edat, tampoc s'observen canvis a tenir en compte.

Passem a analitzar les dades en funció de la xarxa social i la localització del públic. A la taula inferior (Figura 49) es pot observar com la zona *Coast* és la que millor respon a la publicitat per a totes les xarxes socials, mentre que Instagram i Youtube segueixen sent les millors en quant a rendiment per una única zona. Si afegim més detall de localització, la tendència és la mateixa, però veiem que dins d'una mateixa zona geogràfica hi ha ciutats on el rendiment es clarament molt superior a les altres ciutats de la zona. Un exemple clar

es la zona *Centre*, on *Madrid* destaca molt clarament per sobre de les altres ciutats de la seva mateixa zona:

Xarxa social	facebook	instagram	twitter	youtube
Zona	CTR	CTR	CTR	CTR
Centre	9,69%	9,92%	9,69%	9,94%
Coast	10,76%	10,97%	10,71%	11,07%
North	9,68%	9,94%	9,71%	9,99%
South	9,13%	9,34%	9,12%	9,31%

Xarxa social		facebook	instagram	twitter	youtube
Zona	Ciutat	CTR	CTR	CTR	CTR
Centre	Madrid	11,24%	11,49%	11,18%	11,51%
	Toledo	7,58%	7,72%	7,59%	7,81%
	Valladolid	7,24%	7,45%	7,38%	7,42%
Coast	Barcelona	11,22%	11,41%	11,13%	11,52%
	Tarragona	10,15%	10,40%	10,16%	10,46%
	Valencia	10,31%	10,55%	10,29%	10,67%
North	Bilbao	10,82%	11,08%	10,81%	11,17%
	Gijon	8,45%	8,67%	8,48%	8,65%
	La Coruña	8,34%	8,62%	8,44%	8,67%
South	Cadiz	7,92%	8,10%	7,91%	8,15%
	Malaga	9,04%	9,32%	8,98%	9,24%
	Sevilla	9,69%	9,91%	9,72%	9,85%

Figura 49: Taules d'anàlisi en funció de zones-ciutats i xarxa social

Analitzarem com es comporta el rendiment si afegim la dimensió edat a la taula d'anàlisi (Figura 50). La conclusió inicial que es pot extreure és que les ciutats on suposadament hi ha més densitat de població (*Madrid-Barcelona-Bilbao*) el rang d'edat de 61 a 99 respon millor que en les altres ciutats i especialment bé per a les xarxes Facebook i Tiwitter. El rang d'edat de 46 a 60 també respon millor a la publicitat en aquests xarxes mentre que els altres grups d'edat responen millor als anuncis de Instagram i Youtube.

Xarxa social		facebook				instagram				twitter				youtube			
Edat		18-30	31-45	46-60	61-99	18-30	31-45	46-60	61-99	18-30	31-45	46-60	61-99	18-30	31-45	46-60	61-99
Zona	Ciutat	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
Centre	Madrid	10,52%	11,44%	12,62%	10,37%	12,61%	12,53%	11,44%	9,39%	10,48%	11,59%	12,38%	10,26%	12,64%	12,54%	11,58%	9,27%
	Toledo	7,10%	7,70%	8,55%	6,96%	8,56%	8,65%	7,58%	6,08%	6,98%	7,76%	8,53%	7,09%	8,44%	8,69%	7,82%	6,31%
	Valladolid	6,65%	7,39%	8,16%	6,74%	8,26%	8,17%	7,44%	5,98%	6,64%	7,70%	8,37%	6,81%	8,13%	8,06%	7,55%	5,93%
Coast	Barcelona	10,31%	11,59%	12,52%	10,44%	12,32%	12,54%	11,54%	9,23%	10,26%	11,48%	12,50%	10,30%	12,63%	12,61%	11,53%	9,30%
	Tarragona	9,35%	10,32%	11,64%	9,32%	11,58%	11,40%	10,24%	8,41%	9,48%	10,39%	11,36%	9,41%	11,46%	11,30%	10,51%	8,54%
	Valencia	9,72%	10,54%	11,55%	9,46%	11,63%	11,58%	10,51%	8,51%	9,37%	10,49%	11,70%	9,57%	11,66%	11,70%	10,66%	8,68%
North	Bilbao	9,95%	11,15%	12,23%	9,93%	12,17%	12,04%	11,09%	9,01%	9,89%	11,07%	12,16%	10,10%	12,24%	12,12%	11,07%	9,25%
	Gijon	7,70%	8,60%	9,72%	7,79%	9,46%	9,60%	8,68%	6,92%	7,82%	8,70%	9,52%	7,85%	9,55%	9,42%	8,67%	6,98%
	La Coruña	7,74%	8,45%	9,32%	7,87%	9,62%	9,51%	8,45%	6,92%	7,86%	8,60%	9,53%	7,80%	9,52%	9,57%	8,65%	6,97%
South	Cadiz	7,22%	8,18%	8,95%	7,36%	8,86%	8,79%	8,19%	6,54%	7,39%	8,11%	8,89%	7,24%	8,98%	8,86%	8,17%	6,60%
	Malaga	8,46%	9,16%	10,29%	8,27%	10,19%	10,20%	9,35%	7,53%	8,29%	9,22%	10,09%	8,32%	10,06%	10,17%	9,34%	7,39%
	Sevilla	8,99%	9,82%	10,90%	9,07%	10,89%	10,77%	9,96%	8,00%	9,06%	9,95%	10,85%	9,00%	10,87%	10,71%	9,70%	8,14%

Figura 50: Taula d'anàlisi en funció de zones, ciutats, edat i xarxa social

Fins aquest punt s'ha analitzat la resposta del CTR en funció de paràmetres no temporals, fixant-nos en quina o quines relacions podem establir entre els valors del CTR i les dimensions. A continuació s'analitzarà el rendiment de la publicitat al llarg del temps del que es disposen dades (l'any 20017). Si realitzem un anàlisi global per xarxa social, i període temporal (fins arribar al detall dels mesos) s'observa que durant el segon semestre el rendiment de la publicitat decau de forma clara. Si analitzem el que succeeix trimestralment, la tendència es ratifica, i analitzant mensualment, veiem com que a partir de Juny hi ha un descens clar del rendiment (aproximadament un 2%) i que els dos darrers mesos de l'any

es recupera d'una manera molt tímida. En totes les franges temporals s'aprecia que les xarxes socials que millor responen són Instagram i Youtube (com ja s'havia vist anteriorment):

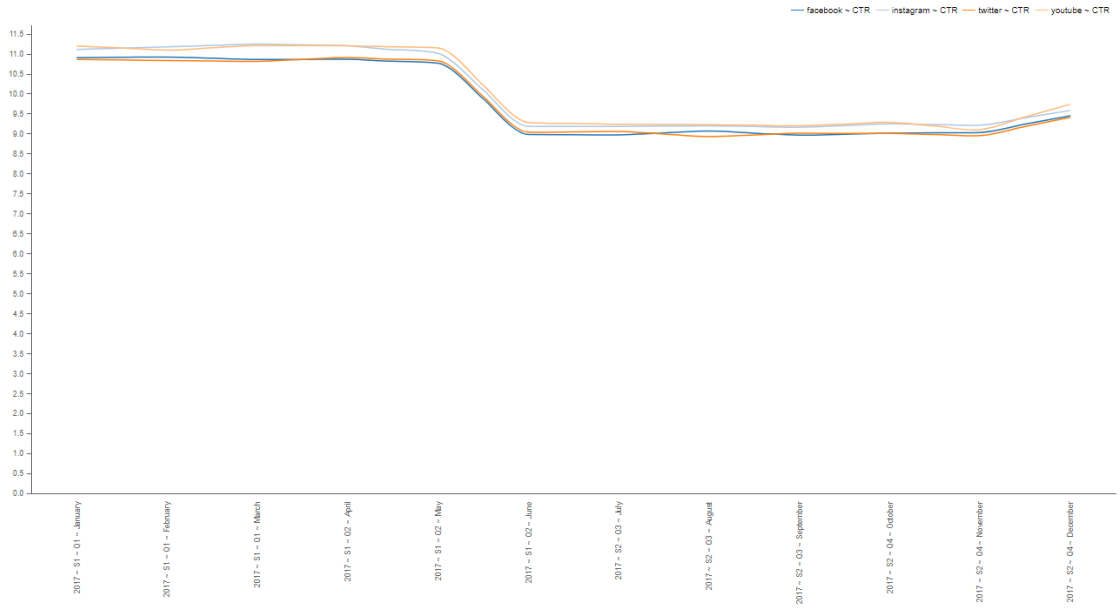
Xarxa social		facebook	instagram	twitter	youtube
Year	Semester	CTR	CTR	CTR	CTR
2017	S1	10,57%	10,84%	10,57%	10,88%
	S2	9,09%	9,27%	9,07%	9,30%

Xarxa social			facebook	instagram	twitter	youtube
Year	Semester	Quarter	CTR	CTR	CTR	CTR
2017	S1	Q1	10,90%	11,18%	10,84%	11,18%
		Q2	10,23%	10,50%	10,29%	10,58%
	S2	Q3	9,01%	9,19%	9,01%	9,23%
		Q4	9,17%	9,35%	9,13%	9,38%

			Xarxa social	facebook	instagram	twitter	youtube
Year	Semester	Quarter	Month	CTR	CTR	CTR	CTR
2017	S1	Q1	January	10,91%	11,11%	10,87%	11,20%
			February	10,93%	11,18%	10,84%	11,10%
			March	10,86%	11,25%	10,82%	11,22%
		Q2	April	10,87%	11,21%	10,91%	11,21%
			May	10,77%	11,02%	10,83%	11,15%
			June	8,99%	9,19%	9,04%	9,29%
	S2	Q3	July	8,98%	9,19%	9,06%	9,24%
			August	9,08%	9,21%	8,94%	9,23%
			September	8,97%	9,17%	9,02%	9,20%
		Q4	October	9,02%	9,25%	9,01%	9,29%
			November	9,04%	9,22%	8,96%	9,11%
			December	9,46%	9,59%	9,42%	9,75%

Figura 51: Taules d'anàlisi del rendiment per xarxa social per períodes de temps

La tendència explicada anteriorment la podem visualitzar gràficament:



Per finalitzar analitzarem les dades per dia de la setmana. En aquest anàlisi s'observa que els valors de CTR són molt estables per dia de la setmana, però hi ha un lleuger increment del rendiment els caps de setmana a totes les xarxes socials.

	Xarxa social	facebook	instagram	twitter	youtube
Year	DayOfWeek	CTR	CTR	CTR	CTR
2017	Sun	9,92%	10,11%	9,96%	10,24%
	Mon	9,89%	10,05%	9,87%	10,06%
	Tue	9,71%	10,03%	9,77%	10,08%
	Wed	9,83%	10,03%	9,78%	10,01%
	Thu	9,79%	10,06%	9,75%	10,07%
	Fri	9,76%	10,06%	9,78%	10,06%
	Sat	9,95%	10,14%	9,88%	10,19%

Figura 52: Anàlisi de rendiment en funció de dia de la setmana i xarxa social

Com es pot observar, hi ha múltiples combinacions de les dimensions que ens aporten informació útil, com per exemple patrons de rendiment per localitzacions. S'ha pogut comprovar que de forma, a priori, no justificada hi ha un descens del rendiment de la publicitat (a totes les xarxes socials) durant el segon semestre.

6.2.4 Existeixen relacions entre plataformes i franges d'edat d'usuaris que provoquin millors taxes de visualització?

Per obtenir les dades necessàries per analitzar la pregunta creem una consulta de visualització de dades on es mostrarà el valor del CTR en funció de la xarxa social (columnes) i de l'edat dels usuaris (files). En relació a la informació obtinguda, podríem dir que efectivament, existeix aquesta relació entre plataforma i franja d'edat. A continuació es mostren les dades obtingudes:

Xarxa social	facebook	instagram	twitter	youtube
Edat	CTR	CTR	CTR	CTR
18-30	9,11%	11,05%	9,09%	11,08%
31-45	10,05%	11,02%	10,10%	11,02%
46-60	11,09%	10,06%	11,02%	10,10%
61-99	9,10%	8,14%	9,09%	8,21%

Figura 53: Relació entre xarxa social i franges d'edat

El rang d'edat que clarament reacciona pitjor és el que va de 61 a 99 anys. El resultat no sorprèn, ja que estem parlant de plataformes digitals, i aquesta franja d'edat comprèn a un segment que clarament no està acostumat a l'ús d'aquest mitjans i per tant no reacciona de forma favorable generant un CTR baix (notar que el número d'impressions es manté estable en les 4 franges sobre els 6.6 milions d'impressions). Sota el meu punt de vista, aquestes taxes més baixes denoten que per tal de refermar la publicitat en segments més profitosos s'hauria d'evitar mostrar anuncis (en cap xarxa social) a aquest rang d'edat.

Pel que respecta a les altres franges notar que els resultats són similars. Tot i això si que es pot concloure que:

- Per 18-30 i 31-45 les xarxes socials que millor funcionen són Instagram i Youtube.
- Per 46-60 i 61-99 les xarxes socials que millor funciona són Facebook i Twitter.

Els menors de 45 anys tendeixen a respondre millor a la publicitat mostrada a Instagram o Youtube, mentre que es majors de 45 responen millor a la publicitat de Facebook i Twitter. Les franges d'edat d'entre 31-45 i 46-60 responen millor a la publicitat en termes generals (potser per factors socials i econòmics).

Basant-nos en les dades obtingudes i únicament en l'indicador CTR, no realitzaria cap acció correctiva per focalitzar els anuncis de les xarxes socials de forma agressiva, tret d'evitar la franja d'edat de majors de 61 anys.

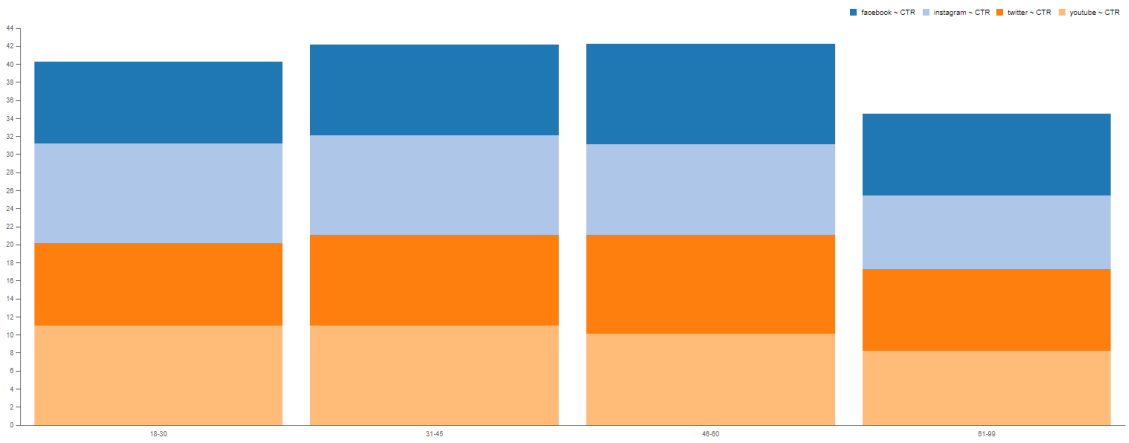


Figura 54:Representació gràfica del rendiment en funció de la localització i producte

En la gràfica s'aprecien les tendències comentades anteriorment, on clarament es pot observar el descens de l'efectivitat en la franja d'edat dels 61 a 99 anys.

6.2.5 El coneixement del grup d'interès dels usuaris podria ajudar a millorar els indicadors per determinats productes?

Per obtenir la informació necessària per poder analitzar la pregunta creem una consulta de visualització de dades on mostrarem el valor del CTR en funció de les aficions (columnes) i de la família de productes (files). En base a les dades que s'obtenen, la resposta a aquesta pregunta és un si rotund. Com és d'esperar, les persones que tenen aficions concretes responen millor (o molt millor) a la publicitat que va segmentada directament a les seves aficions. En el cub analític emprat, es veu molt clarament que el públic amb afició per *Technology* respon molt millor a anuncis de productes electrònics que a anuncis de productes de moda. Això mateix es pot veure amb les aficions *Travel* i *People* que tenen una predisposició més alta envers anuncis de productes vinculats a *Culture*. Un altre exemple el trobem amb productes de la família *Sports* que tenen millor indicador d'efectivitat entre les persones que tenen aficions a *People* i *Sports*. Els resultats obtinguts es mostren a continuació:

Aficions	Business	Cars	Fashion	Garden	People	Sports	Technology	Travel
Familia	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
Accessory	10,75%	15,45%	15,46%	7,79%	12,60%	5,88%	5,91%	5,89%
Culture	7,81%	5,88%	5,90%	7,80%	20,24%	5,90%	5,89%	20,22%
Electronics	5,89%	15,40%	5,90%	5,89%	5,90%	9,70%	24,96%	5,93%
Sports	5,88%	5,89%	5,93%	5,90%	15,48%	24,98%	5,86%	9,70%
Wear	5,88%	10,66%	15,55%	5,88%	7,79%	15,38%	7,82%	10,68%

Figura 55: Taula d'anàlisi de rendiment en funció d'aficions i famílies de productes

En el gràfic de la Figura 52, es pot observar com responen les famílies en relació a les aficions. El gràfic d'estrella destaca clarament la família de productes que funcionen millor per afició (eixos del gràfic). La zona ombrejada de color és correspon amb el valor que pren el CTR per cada afició, quan més a la punta més alt es el valor de CTR i per tant millor és el rendiment del anunci per una afició concreta:

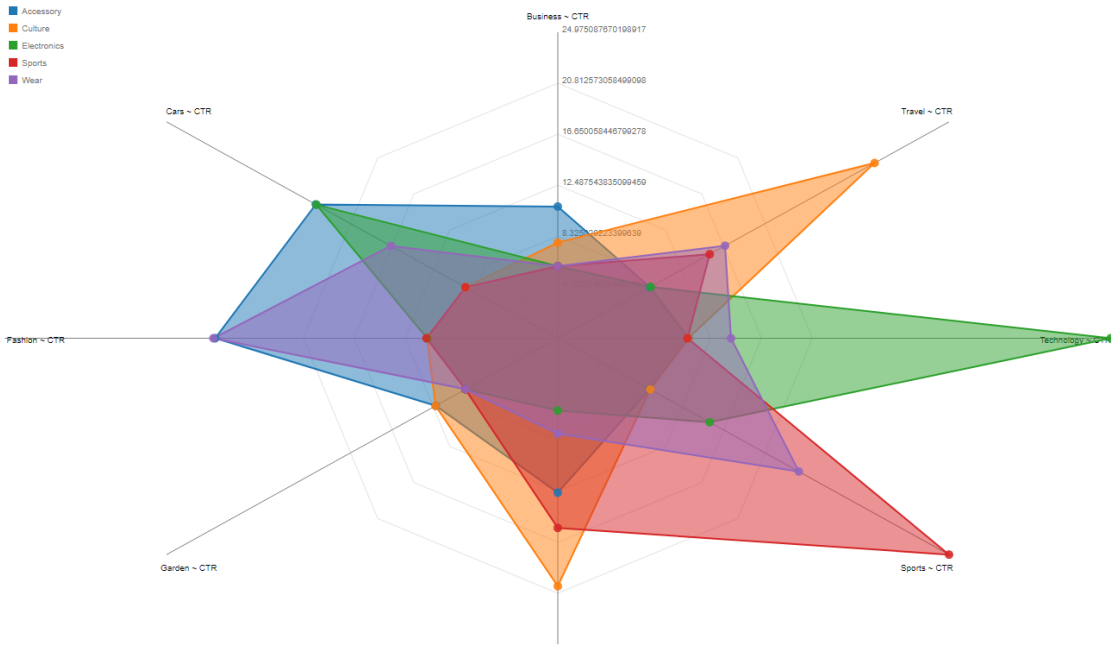


Figura 56: Representació gràfica del rendiment en funció d'aficions i famílies de productes

Una altra representació molt visual ens la mostra el gràfic següent (Figura 53). A l'eix X es representen les aficions, i a l'eix Y les famílies. Quan més gran és el cercle en un X-Y (Afició-Família) més gran és el valor del CTR i per tant millor rendiment tenen els anuncis d'una família determinada per un segment d'aficions.

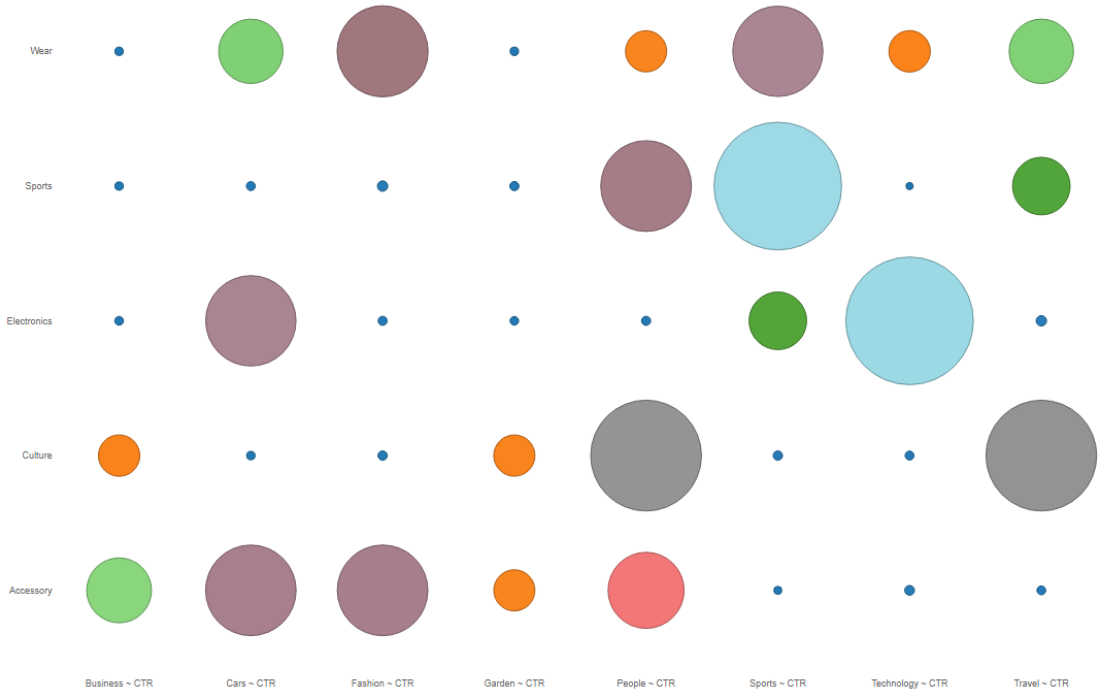


Figura 57:Representació gràfica del rendiment en funció d'aficions i famílies de productes (II)

Continuem amb l'anàlisi i avancem un pas mes enllà, analitzant la resposta envers els productes anunciats. Analitzem si per a productes, aquesta vinculació es segueix mantenint o si per el contrari es perd. Per això modifiquem la visualització de dades anterior afegint un nivell més de detall a la dimensió de productes. A més per facilitar la llegibilitat de les dades obtingudes s'intercanvien els eixos de la taula d'anàlisi. A continuació es mostren les dades obtingudes:

Familia	Accessory		Culture		Electronics	Sports	Wear	
Producte	Scarf	Watch	Theater	Trip	Mobile Phone	Sneakers	Dress	Sheatshirt
Aficions	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
Business	5,90%	15,48%	9,69%	5,89%	5,89%	5,88%	5,88%	5,88%
Cars	5,89%	25,04%	5,88%	5,89%	15,40%	5,89%	5,91%	15,47%
Fashion	25,04%	5,89%	5,89%	5,91%	5,90%	5,93%	25,09%	5,89%
Garden	9,70%	5,89%	5,88%	9,71%	5,89%	5,90%	5,88%	5,89%
People	15,44%	9,72%	25,05%	15,47%	5,90%	15,48%	9,70%	5,89%
Sports	5,85%	5,90%	5,89%	5,91%	9,70%	24,98%	5,88%	25,02%
Technology	5,94%	5,89%	5,91%	5,87%	24,96%	5,86%	5,91%	9,72%
Travel	5,90%	5,89%	15,46%	24,96%	5,93%	9,70%	15,44%	5,89%

Figura 58: Taula d'anàlisi de rendiment en funció d'aficions i productes

Com es podia esperar el vincle persisteix, però la possibilitat d'afegir un nivell de detall major ens mostra dades interessants. La tendència comentada anteriorment es segueix reflectint a nivell de producte. Es destacable el valor d'efectivitat d'algunes combinacions com *Fashion-Scarf*, *People-Theater* o *Fashion-Dress*. Aquests valors d'efectivitat tant alts (prop del 25%) ens indiquen que es molt positiu segmentar els anuncis en funció de les aficions

del públic objectiu, ja que adreçar l'anunci del producte correcte al públic correcte augmenta el rendiment de forma molt clara.

A continuació es mostra la informació comentada anteriorment en format gràfic:

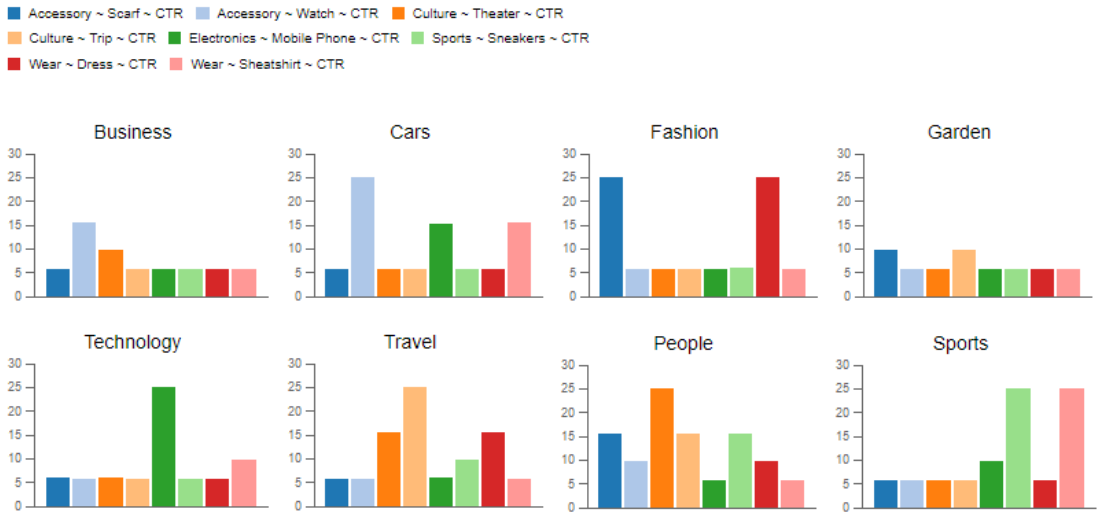


Figura 59: Representació gràfica del rendiment en funció d'aficions i productes

S'observa que el rendiment dels anuncis per a persones amb afició *Garden* és molt inferior a la de la resta de segmentació per aficions. Per aquest motiu es podria proposar de retirar a les persones amb aquesta afició, del públic objectiu objectiu de la plataforma de publicitat online.

7. Conclusions

La realització d'aquest TFM, ha sigut una experiència enriquidora i profitosa en molts aspectes. Personalment, no es tenia cap experiència en la realització de projectes basats en BI, ni es tenia cap tipus de formació o coneixement sobre plataformes BI. L'estructuració del treball, amb la seva concepció per part dels professors de l'àrea de coneixement, i la fragmentació en fases amb objectius concrets i realistes han sigut molt útils alhora d'adquirir els coneixements i competències de forma esgraonada, en el moment precís i a més centrant els esforços en allò que es determina com a important. Serveixin aquestes línies per expressar la meva gratitud a la comunitat educativa de la UOC i en especial a David Amorós Alcaraz (consultor del TFM) i a Maria Isabel Guitart Hormigo (responsable de l'assignatura).

La planificació inicial del projecte s'ha pogut seguir de forma normal, sense més problemes que els esperats per haver d'adquirir coneixements en un àmbit totalment desconegut. La definició de les fites a assolir a cada entrega de les PAC que han composat aquest TFM, ha ajudat a adquirir i consolidar els coneixements necessaris per la seva realització. Algunes de les tasques a realitzar han costat més que d'altres (com és lògic), ja que s'ha de superar una corba d'aprenentatge relacionada amb els conceptes, tècniques aplicades i amb la plataforma a utilitzar. Aquesta corba es supera amb esforç i constància.

Les fases del projecte s'han definit de forma ordenada donant una visió cada cop més profunda dels aspectes clau en un projecte de BI. En relació als objectius assolits els podem resumir en:

- Adquisició de coneixements generals de sistemes BI, des de que són i per a què serveixen fins als components i processos més importants.
- Estructuració d'un projecte BI:
    - Anàlisi del problema a resoldre i definició de l'abast.
    - Disseny del DW a partir del problema que es vol resoldre i les dades que el composen.
    - Definició i implementació dels processos ETL de càrrega del DW. Adquisició de coneixements bàsics de processos ETL.
    - Anàlisi de la informació continguda al DW en relació al problema a resoldre, a partir de cubs OLAP.
- Avaluar de forma objectiva les diferents opcions per realitzar una tasca, per poder realitzar una decisió raonada de quin sistema, tècnica o procediment utilitzar. Aquestes decisions han d'estar raonades en base al propòsit que es vol assolir, fomentant l'esperit crític i justificat de la persona que ha de prendre les decisions.

Com a possibles millores o ampliacions a aquest TFM es podrien explorar les opcions de BI quan els sistemes no són estàtics; amb estàtics es vol expressar que les dades que composen el DW no creixen o canvien al llarg temps. Seria interessant veure com un disseny de sistema de BI ha de fer front a dades que s'actualitzen de forma periòdica en un entorn viu, com és gestionen les càrregues de dades i l'explotació. Aquesta variabilitat del sistema obre les portes a la definició i creació de quadres de comandament per a la visualització de dades i presa de decisions. Opino que podria ser una millora interessant, tot i que s'hauria d'avaluar si els requeriments tècnics per realitzar un projecte d'aquestes característiques el fan viable o no.

8. Glossari

A continuació es mostra una taula amb el resum de termes i abreviatures utilitzades durant el document:

Abreviatura	Descripció
TFM	Treball final de Màster
BI	Business Intelligence
DW	Data Warehouse
CTR	Click Through Rate
ETL	Extracció, Transformació i Càrrega
OLAP	Processament analític en línia (OnLine Analytical Processing)
SCD	Dimensions que canvien lentament (Slowly Changing Dimension)
BD	Base de Dades
SGBD	Sistema de Gestió de Base de Dades
PID	Pentaho Data Integration

9. Bibliografia

[1] https://www.statista.com/statistics/277963/facebooks-quarterly-global-revenue-by-segment/ (2018)

[2] https://www.statista.com/statistics/266249/advertising-revenue-of-google/ (Març 2018)

[3] https://www.statista.com/statistics/449143/twitter-revenue-quarter-segment/ (Marçc 2018)

[4] Ralph Kimball, Margi Ross, The Data Warehouse Toolkit: The definitive Guide to Dimensional Modeling, Tercera edició, Ed. Wiley, Indianapolis, 2013. Pàgina 38.

[5] Ralph Kimball, Margi Ross, The Data Warehouse Toolkit: The definitive Guide to Dimensional Modeling, Tercera edició, Ed. Wiley, Indianapolis, 2013. Pàgina 438.

[6] Ralph Kimball, Margi Ross, The Data Warehouse Toolkit: The definitive Guide to Dimensional Modeling, Tercera edició, Ed. Wiley, Indianapolis, 2013. Pàgina 53.

[7] https://www.pentaho.com/customers (Abril 2018)

[8] https://www.mysql.com (Abril 2018)

[9] https://es.wikipedia.org/wiki/SAP_SE (Abril 2018)

[10] https://es.wikipedia.org/wiki/Oracle_Corporation (Abril 2018)

[11] https://es.wikipedia.org/wiki/IBM (Abril 2018)

[12] https://en.wikipedia.org/wiki/Qlik (Abril 2018)

[13] https://selecthub.com/products/qlikview?from_category=69 (Abril 2018)

[14] https://selecthub.com/products/microsoft-bi?from_category=69 (Abril 2018)

[15] https://es.wikipedia.org/wiki/Microsoft (Abril 2018)

[16] https://opensource.com/business/16/6/top-business-intelligence-reporting-tools (Juny 2016)

[17] https://en.wikipedia.org/wiki/Pentaho (Abril 2018)

[18] https://en.wikipedia.org/wiki/BIRT_Project (Abril 2018)

[19] https://slidex.tips/download/conceptual-modeling-for-etl-process-a-thesis-presented-to-the-faculty-of-califor (2007)

[20] https://www.knowage-suite.com/site/home/ (Abril 2018)

[21] http://www.pentaho.com (Abril 2018)

[22] http://www.eclipse.org/birt/ (Abril 2018)

[23] https://community.jaspersoft.com (Abril 2018)

[24] https://www.knowage-suite.com/site/home/ (Abril 2018)

[25] https://optimalbi.com/blog/2017/02/17/gartner-magic-quadrant-for-business-intelligence-2017-cloud-is-coming-slowly/ (Febrer de 2017)

[26] https://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+Steps (Abril 2018)

[27] María Carina Roldán, Pentaho Data Intgration Beginner's Guide, Segona Edició, Ed. Packt Publishing Ltd, Birmingham 2013.

[28] Introduction OLAP System Components: http://web.mit.edu/profit/PDFS/SlaughterA.pdf (Maig 2018).

[29] https://help.pentaho.com/Documentation/6.0/0N0/020/070 (Maig de 2018)

10. Annexos

10.1 Diagrama de Gantt de la planificació del TFM

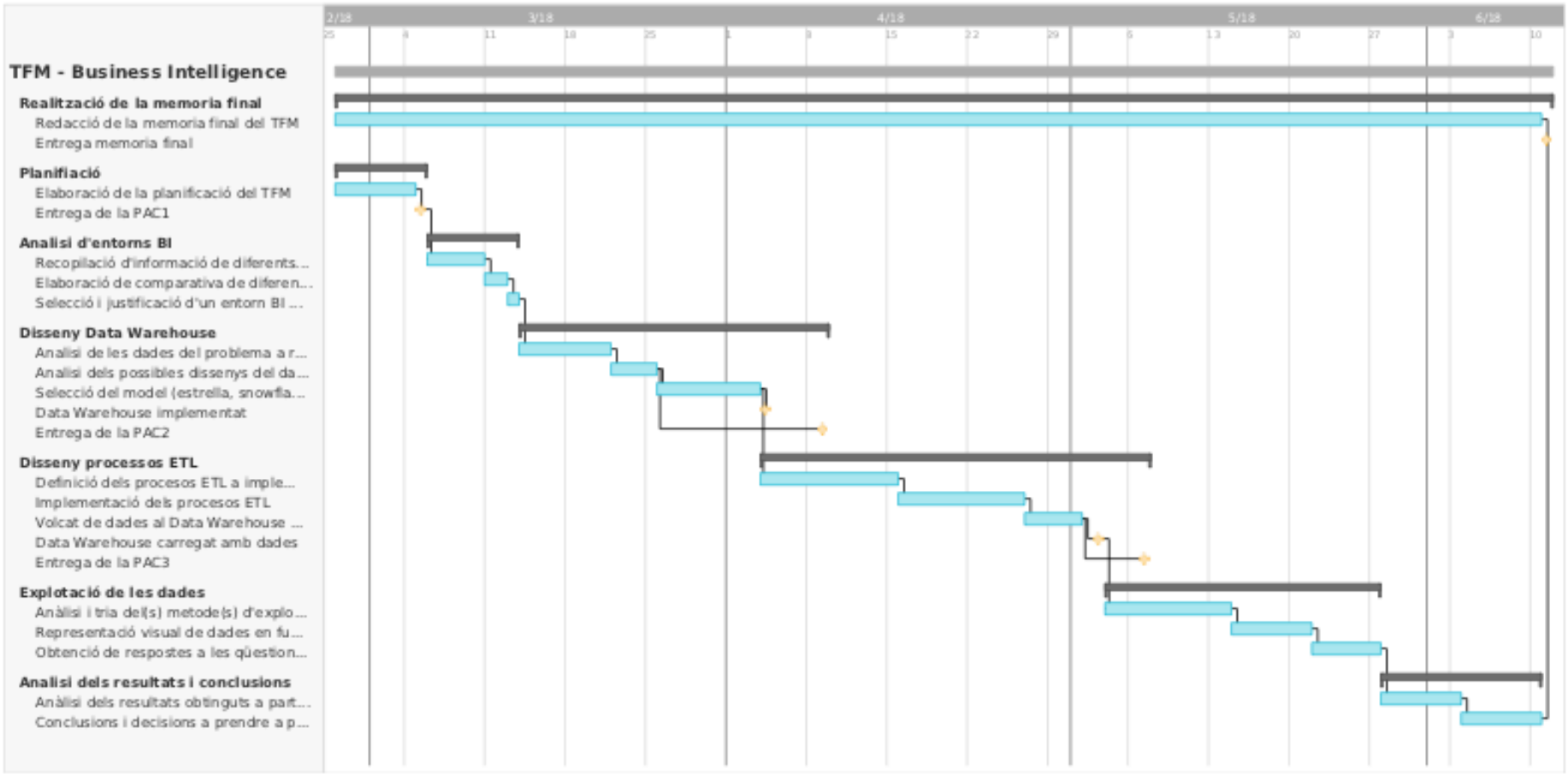


Figura 60: Diagrama de Gantt de la planificació de TFM

10.2 Llista d'elements lliurables que formen part del TFM

Durant la realització d'aquest TFM s'han generat diversos materials que es llisten a continuació i que formen part del recull de productes finals del projecte BI:

1. Script de definició de l'esquema de BD del DW. Capeta "1. Esquema DW", fitxer:
    a. DadesPublicitat.sql
2. Implementació de les transformacions ETL sobre la plataforma Pentaho. Carpeta "2. Transformacions ETL (Pentaho)", fitxers:
    a. FullETLJob.kjb
    b. LoadAgeDimension.ktr
    c. LoadCTRFactChild.ktr
    d. LoadCTRFactParent.ktr
    e. LoadDateDimension.ktr
    f. LoadGenderDimension.ktr
    g. LoadLikeDimension.ktr
    h. LoadLocationDimension.ktr
    i. LoadNetworkDimension.ktr
    j. LoadProductDimension.ktr
3. Fitxer de definició del cub OLAP per Pentaho, utilitzar per realitzar l'anàlisi de dades de la plataforma de publicitat online. Carpeta "3. Cub OLAP", fitxer:
    a. DefincioCubOLAP.xml