Universitat Oberta de Catalunya

uoc.edu

Sistema d'intel·ligència de negoci
Anàlisi de la publicitat en entorns digitals

Francisco José Sánchez Ortuño
Màster Enginyeria Informàtica
Business Intelligence

David Amorós Alcaraz
María Isabel Guitart Hormigo

11 de juny del 2018

FITXA DEL TREBALL FINAL

Títol del treball:	*Sistema d'intel·ligència de negoci. Anàlisi de la publicitat en entorns digitals*
Nom de l'autor:	*Francisco José Sánchez Ortuño*
Nom del consultor/a:	*David Amorós Alcaraz*
Nom del PRA:	*María Isabel Guitart Hormigo*
Data de lliurament:	*06/2018*
Titulació o programa:	*Màster Universitari en Enginyeria Informàtica*
Àrea del Treball Final:	*Business Intelligence*
Idioma del treball:	*Català*
Paraules clau	*Business Intelligence, Publicitat Digital, BigQuery*

Resum del Treball:

La publicitat digital ha evolucionat amb el pas del temps, des de "banners" integrats a pàgines web fins, més recentment, anuncis totalment segmentats dins de xarxes socials. Aquests últims són dirigits per oferir als consumidors continguts del seu interès amb gran precisió i efectivitat. El present treball presenta un sistema d'intel·ligència de negoci o Business Intelligence (BI) que permet analitzar la quantitat d'informació que és generada en campanyes i demostrar l'efectivitat de la segmentació. El sistema implantat s'ha realitzat amb aplicacions de Google Cloud, dins d'un enfocament modern i en el núvol, on les grans quantitats de dades que es poden disposar es puguin processar ràpidament. S'ha utilitzat inicialment una metodologia tradicional però, la desconeixença de la matèria, ha fet que s'hagués d'adaptar a una metodologia més flexible i predisposada als canvis que hi podrien haver. La conclusió d'aquesta implantació és que la integració d'informació de diverses fonts mitjançant eines BI permet una presa de decisions més intuïtiva i molt més àgil, observant les relacions entre productes i segmentacions.

Abstract:

Digital publicity has evolved with the passing of time, from integrated 'banners' on web pages, until more recently, totally segmented advertisements in social networks. These latest advertisements are aimed at offering consumers contents related to their interests with great precision and effectivity. The present work project presents a system of Business Intelligence which allows the analyzation of the quantity of information that is generated in campaigns and demonstrate the effect of follow ups. The implanted system has taken place with applications from Google Cloud, in a modern focus and in a cloud

where the great quantities of data available can be processed rapidly. Initially a traditional method was used but the lack of knowledge of the material has led to the adaption to a more flexible methodology and one that is predisposed to any changes that there could be. The conclusion to this implantation is that the integration of information from various sources through Business Intelligence tools permits the taking of much more intuitive and agile decisions, observing the relations between products and follow ups.

Índex

Llista de figures

1. Introducció

1.1 Context i justificació del Treball

Les xarxes socials han introduït una nova via de comunicació global i, dins d'aquesta, noves formes de fer campanyes publicitàries: la publicitat digital.

Aquestes, a diferència de la publicitat tradicional (anuncis televisius, radiofònics o premsa escrita), no són sistemes monocanals on el que el client que la rep gairebé no hi participa, la comunicació que té és de doble sentit i el client pot participar, generalment donant un "clic" a l'anunci.

Altra diferència fonamental de la publicitat digital és la segmentació o la possibilitat de dividir el mercat al qual es vol dirigir en grups, amb l'objectiu d'augmentar la precisió en l'estratègia que es vol emprar. Més enllà de la segmentació geogràfica que poden donar els mitjans tradicionals, a Internet es pot dividir de forma demogràfica (edat, sexe, nivell d'estudis, etc.) o psicogràfica (personalitat, interessos, gustos, etc.).

Dins dels principals mitjans digitals (Facebook, Youtube, Istagram,...), els anuncis es poden parametritzar segons perfil, localització, edat o aficions focalitzant la seva impressió en un conjunt concret. Proporcionen mètriques estadístiques per mesurar l'efectivitat de les campanyes. Les més importants són Clics i CTR.

Clics: és la quantitat de vegades que els usuaris fan clic a un element publicitari.

CTR (Clic-Through Rate): és la taxa entre el número de vegades que s'ha mostrat l'anunci (impressions) i la quantitat de clics que s'han fet. Donant el percentatge de resposta que té la campanya.

Les plataformes digitals donen les dades i la informació però falta poder extreure el coneixement perquè els administradors publicitaris puguin prendre les decisions més oportunes per tal d'obtenir la millor eficiència.

Els sistemes BI (Business Intelligence) proporcionen el conjunt d'estratègies, dades, aplicacions i tecnologies per aconseguir aquest coneixement. Aquest treball vol realitzar un sistema que permeti veure si hi ha relació entre els diferents paràmetres analitzats i la tipologia de l'anunci. Els resultats obtinguts haurien de poder explicar si la variació en els paràmetres millora els indicadors de Clics i CTR.

1.2 Objectius del Treball

L'objectiu del treball és dissenyar i implementar un sistema d'intel·ligència de negoci que faciliti la presa de decisions a l'hora de triar i administrar campanyes publicitàries en plataformes digitals. Aquest objectiu es detalla en els següents punts:

- Estudi i anàlisi de les diferents eines disponibles en el mercat: que ens aporti les tecnologies i processos per generar una intel·ligència de negoci.
- Selecció per realitzar la implementació entre els productes estudiats: aquell que més s'adapti a les nostres necessitats.

- Disseny i implementació del Data Warehouse: la informació obtinguda de cadascuna de les plataformes digitals serà emmagatzemada per poder ser tractada posteriorment.
- Disseny i implementació dels processos ETL: aquests processos s'encarreguen d'extreure la informació dels diferents orígens, transformar-los a l'estructura del DWH i carregar-los en la base de dades.
- Definició dels procediments per agilitzar consultes: els cubs de processament analític en línia o OLAP permeten superar les limitacions de les bases de dades relacionals, proporcionant una anàlisi de dades ràpida amb una extracció d'informació de les dades.
- Explotació de les dades: les consultes ens han de servir per realitzar informes que donin el coneixement per analitzar i tractar les qüestions del negoci.
- Anàlisi i conclusions: s'analitza totes les parts realitzades del projecte, un cop finalitzat, donant les conclusions i les possibles línies de treball futures.

1.3 Enfocament i mètode seguit

El treball s'ha realitzat en quatre etapes o grups amb l'objectiu d'obtenir els lliuraments necessaris per assolir l'avaluació continua i tenir sempre una idea de la dificultat del projecte. Com no es disposa de dades reals a xarxes socials per poder realitzar les proves, ens hem basat en un full de càlcul amb informació simulada de diferents campanyes publicitàries. Les fases són les següents:

- Pla de treball: un cop s'ha analitzat i s'ha entès la problemàtica del treball, es realitza un pla amb Microsoft Project indicant cada tasca detallada temporalment i les fites que es volen assolir. Per aquest pla es realitza un seguiment; mantenint i revisant que el projecte es dugui a terme correctament i en els terminis previstos.
- Estudi de les plataformes disponibles en el mercat segons objectius del projecte: s'analitza la problemàtica de l'enunciat més en detall per enfocar la recerca de les eines d'intel·ligència de negoci en productes que s'adaptin més a les nostres necessitats. Aquest enfocament pretén que els productes utilitzats siguin fàcils d'implementar per evitar demores en la seva creació i posada en marxa.
- Implementació: es crea el sistema que resolgui els problemes plantejats. La implementació es divideix en tres parts, cada part que componen la plataforma, explicant el seu desenvolupament i anotant les dificultats trobades.
- Finalització de la memòria i presentació virtual: l'edició de la memòria es realitza paral·lelament a la resta de fases, deixant pel final la seva revisió i l'escriptura de les conclusions apreses.

A part de les fases pròpies del treball s'ha fet molta feina d'estudi i aprenentatge de l'àrea de Business Intelligence pel desconeixement de com abordar un projecte d'aquesta temàtica.

1.4 Planificació del Treball

El treball es realitzarà utilitzant eines dins del mercat de intel·ligència de negoci. La primera part del projecte, evidentment, es centrarà en l'estudi i selecció dels productes que ens permeti poder desenvolupar el sistema.

Posteriorment es realitzarà el disseny i desenvolupament de cada una de les parts fins, finalment, trobar les conclusions i possibles millores futures.

El conjunt de tasques i fites es mostren a continuació dins d'una taula i un diagrama de gantt.

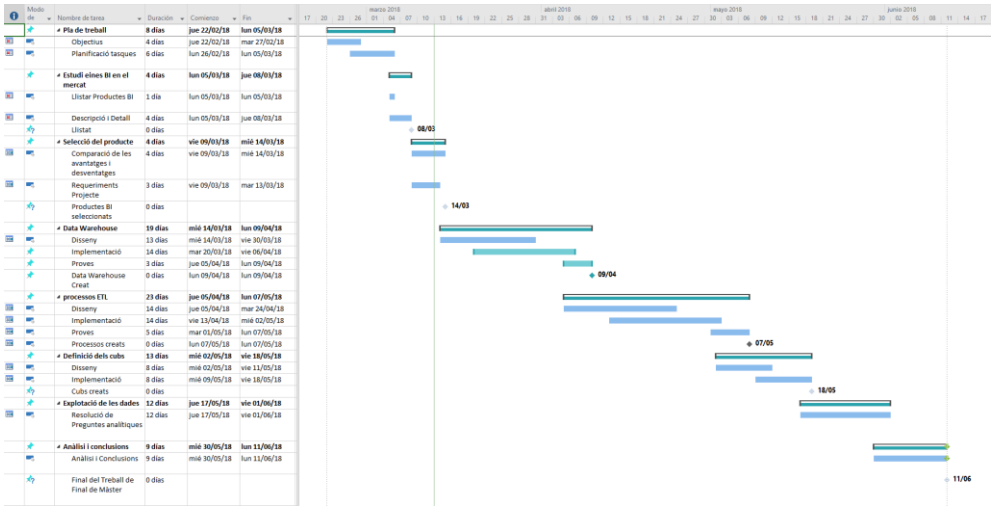


Figura 1: Planificació del treball

1.5 Breu sumari de productes obtinguts

- Memòria del treball realitzat.
- Annexos con procediments d'implementació o instal·lació (dins de la memòria).
- Enllaç amb el quadre de comandament utilitzat per l'analítica.

1.6 Breu descripció dels altres capítols de la memòria

Aquest treball està organitzat de la següent forma. En el capítol 2 expliquem breument que és la intel·ligència de negoci, les seves arquitectures típiques i la que farem servir, s'analitza les diferents eines que trobem en el mercat i indiquem les triades. En el capítol 3 expliquem com és el magatzem de dades, el disseny del seu model i com s'ha implementat. El capítol 4 tractem els processos d'extracció, transformació i càrrega de les dades, s'explica un model teòric de com serien aquests processos per un projecte real i un altre pràctic i implementat utilitzant les dades simulades disponibles. El capítol 5 expliquem com consultar la informació emmagatzemada d'una forma la més eficient possible i contem perquè no es fa un ús de tècniques com els cubs OLAPs (típics en sistemes de Business Intelligence). En el capítol 6 es responen a totes les preguntes analítiques que s'han plantejat en el treball final de màster i s'explica l'eina utilitzada per trobar la resposta. Per finalitzar, es comenten les conclusions i reflexions del treball.

2. Arquitectura de la plataforma BI i estudi d'eines de mercat

Quan parlem d'intel·ligència de negoci o BI (per les seves sigles en anglès) ens referim al conjunt de processos que transformen les dades d'una companyia en informació i coneixement per donar suport a la presa de decisions. Les dades són uns actius d'una organització que, si s'aprofiten al màxim, podem obtenir un avantatge competitiu, com augmentar ingressos o reduir costos, sempre si són analitzades correctament.

En l'àrea del Business Intelligence trobem una varietat d'iniciatives per a aplicar aquests processos, on hi ha una tendència a uns sistemes moderns però amb diferents opcions de tecnologia i implementació, cadascuna amb els seus avantatges i inconvenients, i que ens permeten construir un sistema que:

1) Reculli i identifiqui les dades que resideixen en les diverses fonts (les plataformes digitals que volen analitzar)
2) Faci que aquestes dades estiguin disponibles per fer l'informe i l'anàlisi i que ens ajudi a respondre les preguntes del treball.

La plataforma BI a generar pot tenir una arquitectura diferent segons els objectius de negoci i el nivell de modernitat que es vol obtenir, així com la mida i complexitat de les seves operacions o la tecnologia que es disposa. Això fa que la selecció d'eines i aplicacions estigui condicionada al nivell de capacitat perquè porten uns avantatges i limitacions distintes. Una plataforma podria no treballar bé amb grans quantitats de registres i no ser apta per implementar un sistema enfocat al Big Data o les necessitats de l'empresa podrien no ser compatibles amb productes que treballin en el núvol.

A les plataformes tradicionals les tasques tant d'adquisició de dades com el desenvolupament d'informes les realitzen el departament de TI. Els usuaris de negoci no saben on està la informació ni l'estructura o els components del sistema. Quan volen un informe l'han de demanar a l'àrea de TI amb la corresponent demora de temps.

Les plataformes modernes d'anàlisi i BI es caracteritzen per disposar eines fàcils d'usar i que admeten una gamma completa de capacitats analítiques. No requereixen una participació significativa de TI per predefinir els models de dades, com un requisit previ per a l'anàlisi, i en alguns casos, generen automàticament un model de dades reutilitzable. El departament de TI habilita la plataforma BI per tal que la puguin fer servir els usuaris de negoci i aconseguir una eina que cobreixi les seves necessitats d'autoservei.

Per fer els estudis que ens han d'ajudar a crear la nostra plataforma BI ens hem basat, principalment, en diversos articles realitzats per la prestigiosa consultora Gartner. La nostra idea és de poder crear un sistema amb característiques d'un sistema modern.

2.1 Estudi de l'arquitectura de la plataforma BI

La majoria dels projectes BI tenen una estructura comuna amb uns components típics per on flueixen les dades, tal com es mostra al següent diagrama:

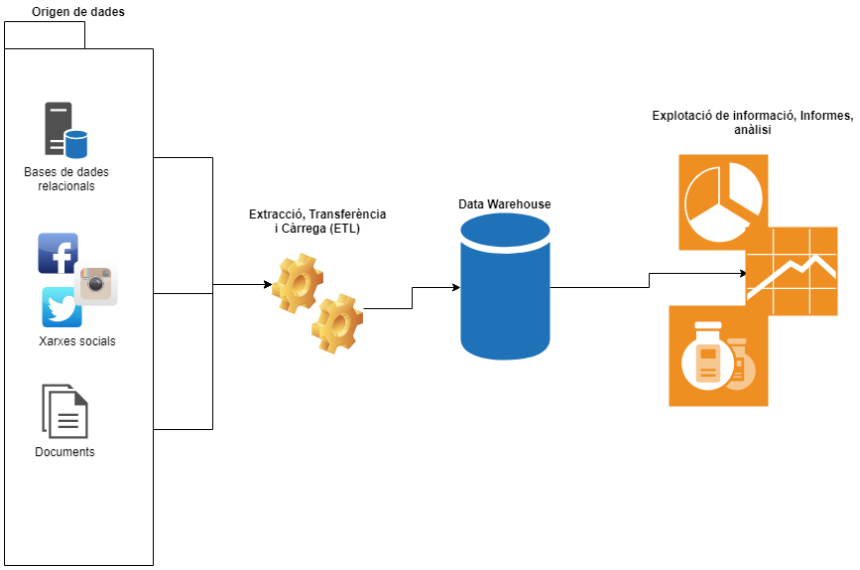


Figura 2: Estructura plataforma BI

El component ETL s'encarrega de l'adquisició de les dades de les diferents fonts (bases de dades relacionals, informació de plataformes digitals, documents d'ofimàtica, etc.). Aquesta tecnologia ha de saber manegar dades de diferents tipus i, a vegades, de qualitat baixa per adaptar-la, netejar-la i carregar-la al magatzem d'on es consultarà la informació unificada.

El magatzem de dades o Data Warehouse conté les dades unificades i permet fer consultes d'una manera eficient i confiable basant-se en un model de dimensions i fets.

L'explotació del magatzem de dades es farà amb eines BI que proporcioni el coneixement al personal de negoci. Si ens fixem en les recomanacions de Gartner, la plataforma BI s'estructura en tres capes, cadascuna amb més funcionalitats analítiques que la de sota:

- Portal d'informació: proporciona informes o quadres de comandament generats pel departament de tecnologia de la informació (TI).
- Banc de treball d'analítiques: les eines BI permeten al personal de negoci generar els informes i analítiques de forma autònoma. El departament de TI habilita aquestes eines o proporciona la infraestructura.
- Laboratori científic de dades: requereix personal especialitzat en tecnologies d'anàlisi i coneixement ampli del negoci, per realitzar models predictius, aprenentatge automàtic o altres capacitats sofisticades.

Depenen del nivell que es vol obtenir, serà més convenient uns productes o uns altres. En els objectius del treball de final de màster no s'inclouen les capacitats d'un laboratori científic. Una estructura que proporcioni informes, o que els permeti generar per tal de resoldre les qüestions del negoci, hauria de ser suficient.

2.2 Estudi d'eines de mercat per la implantació del sistema BI

Per l'anàlisi dels proveïdors, s'ha fet un estudi de les eines que poden donar la funcionalitat per a cadascuna de les capes de l'arquitectura detallada abans (ETL, Data Warehouse i aplicacions BI). Centrant-nos principalment en el magatzem de dades i posteriorment triant les tecnologies que millor s'adaptin.

La instal·lació del magatzem de dades pot ser en una infraestructura local o "On premise", és a dir, utilitzant els recursos propis de l'empresa, o dins d'un servei en el núvol, amb productes que ofereixen proveïdors a Internet.

Els principals avantatges de muntar un sistema local són la latència de les connexions (no ens hauria d'afectar en el nostre cas perquè per mostrar informes i gràfiques no és prioritària) i la por o desconfiança de tenir informació sensible en mans de tercers (tampoc és el nostre cas perquè les dades que volem tractar provenen d'Internet). Quan no tenim el sistema local es perd control del sistema perquè es depèn del proveïdor del servei.

Tenir la base de dades al núvol, al contrari, ens dóna una escalabilitat de la infraestructura que no pot donar cap sistema local, la memòria i/o capacitat de càlcul augmentarà en funció de les nostres necessitats. Aquesta és gestionada pel proveïdor que ens estalvia les tasques de manteniment i incidències del maquinari, donant una flexibilitat i agilitat a l'hora de crear el Data Warehouse. Pel nostre projecte l'escalabilitat no ens suposa un avantatge (per la quantitat de dades que maneguem inicialment) però sí que ens interessa aquesta agilitat i estalvi de recursos propis que ens dóna el núvol. Per això, per fer l'estudi, ens centrarem en aquestes solucions que poden ser instal·lades a Internet.

Els serveis en el núvol es poden adoptar de distintes formes, entre elles es troben: Infraestructura com a servei (IaaS, Infrastructure as a Service), Plataforma com a servei (PaaS, Platform as a Service) o Programari com a servei (SaaS, Software as a Service).

Cada model ens dóna un nivell diferent de gestió. En un IaaS el proveïdor extern ens proporciona tota la infraestructura de servidors virtuals i connexions a xarxes, nosaltres hem de fer la instal·lació de la plataforma (sistema operatiu o entorn) i el programari de la base de dades. A PaaS, a més de la infraestructura, tenim l'entorn on només ens hem de preocupar de fer la construcció i posada en marxa de la base de dades. I per últim, a un model SaaS, disposem directament del programari on, amb opcions de gestió proporcionades pel proveïdor, podem administrar la base de dades sense haver de mantenir ni maquinari ni la plataforma on es troben les nostres dades.

En la recerca de l'eina s'inclourà productes de bases de dades com a servei com programaris on es pugui implementar dins d'un model PaaS.

Els quadrants màgics de Gartner ens permeten veure quins productes o solucions dels diferents fabricants es troben en el lideratge, o quins són els proveïdors millors pel meu projecte. De forma gràfica poden veure els proveïdors líders, els aspirants, visionaris o jugadors de nínxol.

El quadrant de la imatge (Gartner Inc. Magic Quadrant for Data Management Solutions for Analytics, 2018) és l'estudi de les solucions d'administració de dades per a l'anàlisi (DMSA), sistemes de programari complet per l'administració de dades optimitzades específicament per donar suport al processament analític.

Figura 3: Quadrant màgic de solucions d'administració de dades per a l'anàlisi

D'aquí podem extreure quina eina s'ajusta més als nostres requisits perquè ens ajudi en la creació del nostre Data Warehouse. Hem preseleccionat les següents solucions.

Proveïdor	Solució	Virtuts	Inconvenients
Amazon Web Services	Amazon Redshift. L'empresa de Seattle, Washington, EE. UU, ofereix un servei PaaS d'emmagatzematge en el núvol, senzill de fer servir i totalment administrat. S'integra amb eines ETL i BI externes per analitzar les dades o mitjançant Amazon Spectrum, que optimitza les consultes en Amazon S3	Domini en el núvol. Els serveis AWS són regnants amb un marge únicament proper per Microsoft i poden ser utilitzats per a diferents casos d'ús del DMSA. Té preus a baix cost i sota demanda depenen dels recursos que es vulguin crear.	Només ofereix servei en el núvol. No es pot fer una instal·lació en la mateixa empresa (on-premise) o híbrida. L'escalat de recursos no és independent a Amazon S3, té una configuració fixa. Integració dels serveis complexa.
Google	Amb seu en Montain View, California, l'empresa subsidiària del holding Alphabet disposa de diverses ofertes de dbPaaS (Plataforma de Base de dades com a servei), entre elles esta BigQuery, un magatzem de dades administrat enfocat a la informació empresarial o BigTable, una base de dades NoSQL amb baixa latència i alt rendiment en carregues de treball. En el quadrant de Gartner no està com a líder però l'hem preseleccionat per la seva evolució i expectatives futures	Arquitectura moderna en el núvol i preus. Google ha invertit molt en la seva plataforma aprofitant una xarxa d'alta velocitat, oferint productes enfocats a sense servidor (serverless) implementats internament durant anys. Facilitat d'implementació i relació qualitat-preu. Evolució i avanços importants en el mercat de DMSA posicionant-la com una empresa destacable en el futur	Suport i documentació de la plataforma encara no és suficientment madur. Pocs proveïdors de tercers pel desenvolupament de l'ecosistema, tanmateix, Google ha pres mesures i s'espera que millori la situació
Microsoft	L'empresa de Redmond, Washington, disposa de solucions locals com Analytics Platform System o en el núvol com SQL Data Warehouse i que proporciona processament paral·lel massiu (MPP). També tenen Azure HDInsight, servei en núvol administrat amb marcs de codi obert com Hadoop, Spark o Kafka que permet el processament de grans quantitats de dades	Catàleg de productes locals o híbrids amb una integració perfecta amb productes en el núvol. Bona relació qualitat-preu	Atenció al client i manca de suport. Escalabilitat de lectura inferior a altres competidors
Oracle	Oracle, amb seu en Redwood Shores, California, té una cartera de bases de dades com Oracle Database 12c, Oracle Autonomous Data Warehouse o Oracle Exadata que permeten implementar-se tant en centre de dades del client com a centres en el núvol d'Oracle	Líder durant anys en tecnologies DBMS i alta presència en el mercat	Preus alts i capacitats en el núvol no provades

A continuació es fa la recerca dels productes ETL que alimentin una de les solucions triades anteriorment. Aquests han de ser capaços de connectar-se a les bases de dades de les xarxes socials, sigui directament o través d'un petit programari, i de replicar la informació en el nostre Data Warehouse destí.

La busca s'ha realitzat mitjançant consultes en buscadors d'Internet i webs especialitzades, prioritzant que siguin fàcils d'utilitzar i de connectar-se o integrar-se a les diferents fonts de dades i magatzems preseleccionats. També ens ha ajudat el quadrant de Gartner d' Eines d'integració de dades (Gartner, Magic Quadrant for Data Integration Tools, 2017). La busca dona els següents resultats:

Solució	Descripció	Integració	Inconvenients
Talend Open Studio	Programari de codi obert, proporciona més de 800 connectors (més que cap altre), fent que sigui molt fàcil implementació amb diverses bases de dades, formats de fitxers o aplicacions empresarials. Conté eines gràfiques i assistents que simplifica la generació de codi.	Destins de les dades als sistemes d'Amazon Redshift, Google BigQuery i Azure HDInsight	Problemes d'estabilitat en les noves versions.
Jaspersoft ETL	Fàcil de fer servir, extreu les dades de sistemes transaccionals per crear un magatzem consolidat o una plataforma d'anàlisi. Conté altres solucions per generar la plataforma d'intel·ligència de negoci.	Destins de les dades als sistemes de Google BigQuery i Amazon Redshift.	És una subdivisió de Talend, més fàcil de fer servir però amb algunes limitacions respecte a el seu predecessor.
Pentaho Data Integration	Permet als usuaris ingerir, barrejar, netejar i preparar diverses dades de qualsevol font. Amb eines visuals per eliminar la codificació i la complexitat, Pentaho posa totes les fonts de dades i les millors dades de qualitat a l'abast dels usuaris de negocis i TI amb implementacions en el núvol, local i hibrida.	S'adapta a Amazon Redshift i permet connexions amb JDBC o ODBC.	Entorn de desenvolupament millorable (segons comentaris de clients) i enfocament a les solucions de grans quantitats de dades.
Stitchdata	Programari propietari que permet omplir de dades el Data Warehouse en minuts sense manteniments d'APIs, treballs CRON, scripts o JSON.	Integració amb Facebook Ads i destinació de les dades en Amazon RedShift i Google BigQuery.	Es programari propietari tot i que té una edició gratuïta.
Google Dataflow i Dataprep	Dataflow és un model de programació i un servei per desenvolupar i executar patrons de processament, entre ells ETLs. El servei és administrat i gestionat de forma transparent. Dataprep és altre servei de Google desenvolupat juntament amb Trifacta pel tractament i neteja de dades. Permet de forma gràfica i fàcil crear fluxos per a Dataflow	S'integra la resta de serveis de Google com Cloud Storage, Bigquery o BigTable. També té altres integracions mitjançant socis i desenvolupadors externs.	Dataprep està només en versió Beta.

Per últim es fa l'estudi de les eines pròpies de BI que trobem al mercat. Ens hem basat en recerques de productes per plataformes modernes, tal i com s'ha explicat anteriorment, enfocades en l'agilitat i autonomia dels informes. També ens hem recolzat amb un quadrant de Gartner per a plataformes d'Analytics i Business Intelligence (Gartner, Magic Quadrant for Analytics and Business Intelligence Platforms, 2018).

Els criteris que s'ha seguit són: connectivitat amb els magatzems preseleccionats; capacitat de gestió de les metadades perquè es pugui buscar, capturar i reutilitzar dimensions, jerarquies, mesures, etc. amb autoservei; capacitat de donar la informació de forma visual mitjançant gràfics i amb opció d'interactuar amb ells; Facilitat d'ús, atractiu visual i integració amb el flux de treball. La taula següent dona els resultats de l'estudi:

Solució	Descripció	Integració	Inconvenients
Tableau	Programari líder per visualitzar informació de forma gràfica i interactiva sense la necessitat de tenir coneixements tècnics o de codificació. Disposa de 3 productes comercials (escriptori, servidor i núvol) i dos gratuïts (mòbil i públic). Té l'opció d'implementar-se localment o en el núvol, oferint màquines virtuals per AWS, Azure o Google Cloud.	Té connexions natives amb Amazon Redshift, Google Bigquery i Azure HDInsight o SQL Data Warehouse entre moltes altres.	Preus elevats per llicències en comparació a altres competidors. Suport de dades complexes formades per grans i variades fonts d'informació.
Qlik	Ofereix anàlisi de dades àgils i governats amb l'aplicació Qlik Sense. Analítica integrada i preparada amb Qlik Analytics Platform. I creació de quadres de comandament i descobriment de dades amb QlikView. Els seus productes són escalables i preparats per a grans quantitats de dades complexes. Socis externs poden ampliar la plataforma desenvolupant contingut nou, el 70% de les implementacions de Qlik són externes.	Dins del mercat de desenvolupadors trobem connectors de pràcticament qualsevol font de dades: Amazon Redshift, Google Bigquery, Oracle Essbase i connectors amb ODBC.	Preus i dificultat perquè un usuari no tècnic tingui autoservei per dissenyar analítiques.
Google Datastudio	Eina de Google per visualitzar dades i crear quadres de comandament. Permet realitzar una anàlisi de la informació d'una manera molt fàcil. S'encarrega de l'autenticació, drets d'accés i estructuració de les dades. Transforma les dades en mètriques i dimensions.	Disposa de connectors pels diferents productes de Google Cloud, entre ells Bigquery. També dona la possibilitat de crear nous connectors o d'afegir de creats per la comunitat de desenvolupadors	Es troba en fase Beta i no té totes les capacitats que la resta de competidors.
Looker	Plataforma d'intel·ligència comercial (BI), basada en núvol, dissenyada per explorar i analitzar dades. Es pot "veure la font" per comprendre com s'estan manipulant les dades que s'estan visualitzant. Els taulers de comandament permeten presentar dades i informació mitjançant taules, gràfics i informes personalitzables. Tots els quadres de comandament i consultes es poden crear per tal que els usuaris puguin descobrir informació en múltiples nivells.	Connectors de Amazon Redshift, Google Bigquery i Azure HDInsight.	Poca documentació o ajuda i difícil de configurar.

2.3 Eines utilitzades per realitzar el treball

En la selecció de les eines s'ha prioritzat la facilitat d'implementació i els coneixements actuals de les distintes plataformes. Les característiques del projecte fan, que amb qualsevol eina preseleccionada, es pugui realitzar el sistema que ens ajudi a donar informació sobre les diferents campanyes publicitàries.

Molts experts TI en intel·ligència de negoci proposen que no és necessari fer servir les eines d'un únic proveïdor, sinó que segons les necessitats una

aplicació és millor utilitzar els productes de varis. En el nostre cas, però, ens hem decantat en utilitzar les eines de Google.

S'ha triat aquesta opció perquè Google és una empresa confiable, perquè té uns productes amigables de manegar, perquè té una bona documentació i perquè ja es coneixia el seu entorn.

2.3.1 Google Cloud

La plataforma Cloud ens permet desenvolupar aplicacions amb les mateixes infraestructures, eines i tecnologies que utilitza Google pels seus productes, com el buscador d'Internet, correu electrònic (Gmail), Youtube, etc. També inclou programari de les diferents innovacions, sobretot dins de l'àmbit de la intel·ligència artificial i l'aprenentatge automàtic, en forma d'APIs.

Els productes de Google Cloud són tarifats per pago per ús. Aquest ús es mesura en temps, quantitat de memòria, consultes, tràfic, etc. segons el producte, pagant només pel que s'utilitza i amb un límit gratuït que, si no se supera, no s'ha de facturar res. D'aquesta manera permet models de costos variables o predicibles en base de càrrega esperada.

La figura següent mostra l'estructura del projecte detallant els productes utilitzats des de l'adquisició de les dades fins a la visualització de la informació.

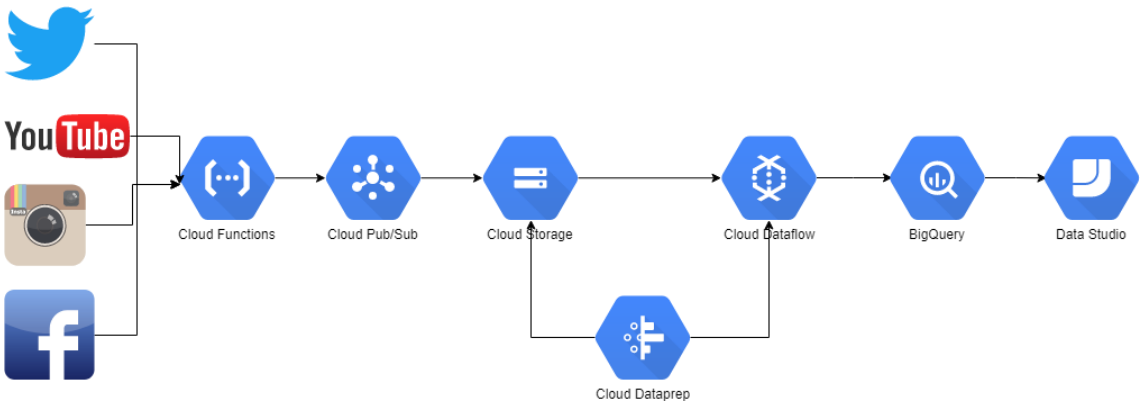


Figura 4: Estructura del sistema BI de Google

2.3.2 BigQuery

Cloud BigQuery és la tecnologia recomanada de Google per implementar el magatzem de dades per a intel·ligència empresarial.

La seva funció és d'unificar les dades de les diferents fonts per poder-les consultar i extreure la seva informació.

2.3.3 Cloud Functions i Cloud Pub/Sub

Cloud Functions permet crear petites aplicacions sense servidor. Aquestes es poden executar sota demanda segons les necessitats del moment. Cloud Pub/Sub és un servei de recepció i enviament de missatges, útils per serveis en entorns "Streaming". Permet sincronització asíncrona i rendiment i latència constants.

Aquests dos productes ens ajudaran a crear la connexió de cada plataforma digital i extreure les dades.

2.3.4 Cloud Dataflow i Cloud Dataprep

Google Dataflow s'integra amb Pub/Sub per processar les dades, generant les transformacions i neteges necessàries per poder-se incorporar a Bigquery. Dataprep ens permet analitzar i fer les transformacions d'una manera gràfica i així agilitzar les tasques de crear Dataflow. Per fer les proves agafarem les dades allotjades en Cloud Storage i incorporades de l'arxiu Excel simulat.

2.3.5 Data Studio

Data Studio serà el nostre sistema de consulta i visualització. Es crearà diferents quadres de comandament amb connexions a Bigquery.

3. Data Warehouse

El magatzem de dades o Data Warehouse és un element clau en la intel·ligència de negoci (tot i que podria ser prescindible). Emmagatzema les dades i la informació que l'empresa pot utilitzar per fer l'analítica i la pressa de decisions. Normalment les dades provenen de diverses fonts amb l'objectiu d'unificar la informació i així trobar relacions entre elles.

3.1 Característiques

El producte que ofereix Google té un enfocament modern, administrat, sense servidors, dissenyat per fer consultes SQL ultraràpides i de baix cost.

És modern perquè és àgil; fàcil de fer servir; i de confiança per donar suport al rol d'impulsar la competitivitat i de crear valor en el negoci. Dissenyat per suportar el ritme actual on, el canvi i el creixement de fonts, volum i complexitat de les dades augmenten exponencialment.

És administrat perquè no s'han de crear servidors, ni clústers, ni sistemes operatius i no s'ha de fer el manteniment de maquinari.

És un servei en el núvol i per tant té els inconvenients d'aquests serveis: hi ha una dependència amb la connexió a Internet i les dades poden estar emmagatzemades en un país que tingui una legislació de protecció de dades no reconeguda per la Unió Europea (no és el cas de BigQuery en quant a protecció de dades). Però també té els seus avantatges: accés des de qualsevol lloc amb Internet i per mitjà de qualsevol dispositiu; Recursos agrupats i elàstics, Google dóna una capacitat de càlcul que s'adapta segons les necessitats de les consultes i que es pot ampliar en casos especials; Servei mesurat, tenim diverses mètriques per controlar i optimitzar el sistema.

Està pensat per fer consultes de grans quantitats de dades en pocs segons. BigQuery és la capa externa de Dremel (Google, Inc., 2010), tecnologia que combina arbres d'execució de diferents nivells i dades en columnes. Escala a milers de CPUs i petabytes d'informació per fer una computació basada en MapReduce. Les consultes s'escriuen en SQL estalviant-nos la costosa programació de MapReduce.

Google es caracteritza per oferir serveis de baix cost. Només cobren per l'ús del sistema, és a dir, per la computació de les consultes i per l'espai ocupat.

3.2 Disseny del model de dades

Els models de dades dels Data Warehouse fan servir una tècnica de modelatge dimensional que siguin compatibles amb un entorn d'anàlisi i que permetin realitzar consultes de manera més òptima. Farem servir l'enfocament de Ralph Kimball (Kimball, 2013) orientat a la consulta d'informació.

Per realitzar el modelatge s'ha d'entendre i analitzar els següents conceptes:

- Normalització o desnormalització: que la base de dades estigui normalitzada farà que n'hi hagi menys redundància (els registres ocuparan menys espai) i les actualitzacions seran menys costoses (model típic de les bases de dades relacionals). En canvi, en una desnormalitzada s'aconseguirà consultes més ràpides en estalviar-nos fer unions entre taules però la repetició d'informació fa que s'ocupi més espai en memòria.
- Fets i dimensions: quins valors són indicadors de negoci (Fets) i sobre quins atributs es voldran analitzar (dimensions).
- Esquema en estrella o floc de neu: si les dimensions s'implementen en vàries taules per tal de normalitzar la informació farà falta un esquema de floc de neu.
- Granulalitat o jerarquies: quin nivell de detall es vol, tant pels fets com de les dimensions.

Les dades que farem servir són les simulades en format Excel, es crea inicialment un model de dades lògiques semàntiques en forma d'estrella (seguint els estàndards de Kimball)

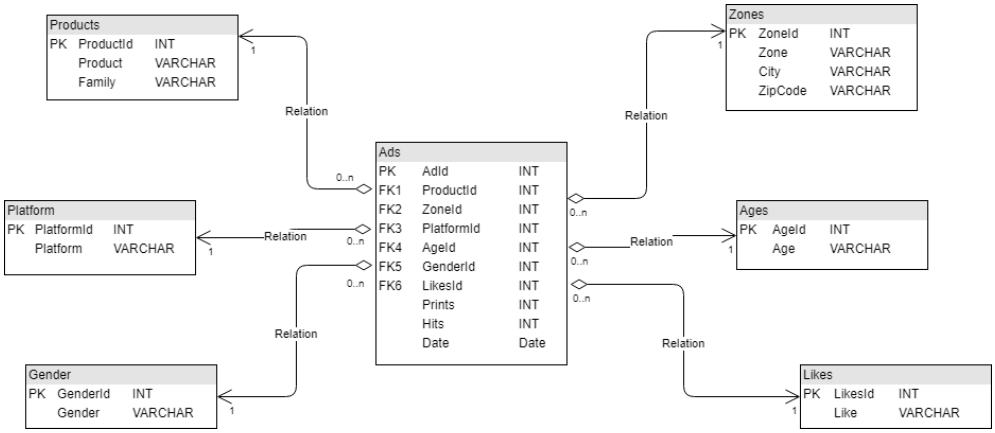


Figura 5: Esquema de la base de dades

Del model estrella podem identificar una taula de fets que representa el procés de negoci que volem representar, en el nostre cas el rendiment dels anuncis. Cada fila representa una campanya d'un anunci concret.

Ads		
Id	INT	Identificador únic
Prints	INT	Número d'impressions de l'anunci
Hits	INT	Número de Clicks fets als anuncis
Date	DATE	Data de l'anunci amb dia, mes i any. Les regles de negoci demana que la informació tingui una granularitat mínima d'un dia i jerarquies de setmanes, mesos, trimestres i anys
ProductId	INT	Identificador de producte. Clau forana
ZoneId	INT	Identificador de zona. Clau forana
PlatformId	INT	Identificador de plataforma. Clau forana
AgeId	INT	Identificador de segmentació. Clau forana
GenderId	INT	Identificador de gènere
LikesId	INT	Identificador de interessos

Les taules de dimensions contenen els atributs descriptius utilitzats per les aplicacions de BI per a la lectura i l'agrupació dels fets. Identifiquem les següents dimensions amb les seves jerarquies:

- Tipologies de articles

Products		
ProductId	INT	Identificador del producte
Product	VARCHAR	Descripció
Family	VARCHAR	Família

- Limitacions geogràfiques on s'han impresos els anuncis

Zones		
ZonaId	INT	Identificador de la zona
Zone	VARCHAR	Zona que engloba la ciutat
City	VARCHAR	Ciutat que engloba el codi postal
ZipCode	VARCHAR	Codi postal

- Plataforma utilitzada per la campanya

Platforms		
PlatformId	INT	Identificador de la plataforma
Platform	VARCHAR	Nom de la plataforma o xarxa social

- Rangs d'edat

Ages		
AgeId	INT	Identificador del rang
Gender	VARCHAR	Rang d'edat a qui va dirigit l'anunci

- Sexe

Genders		
GenderId	INT	Identificador del gènere
Gender	VARCHAR	Sexe a qui va dirigit l'anunci

- Interessos

Likes		
LikeId	INT	Identificador de interès
Like	VARCHAR	Interessos generals que ha de tenir la persona a qui va dirigit l'anunci

3.3 Implementació del model de dades

Per la creació del model en BigQuery, es vol que la informació sigui fàcil d'utilitzar però que també tingui un accés el més eficient possible. Google ven que pot fer consultes de milions de registres en pocs segons però recomana optimitzar la base de dades i el seu esquema per millorar l'eficiència tant en temps com en espai.

BigQuery admet un esquema en forma d'estrella o de floc de neu (normalitzat) però les unions requereixen coordinació de dades i un augment en l'amplada de banda de la comunicació. Les consultes es divideixen en ranures, depenen la seva complexitat, i pot comportar un sobrecost si les dades estan en diferents taules.

Una solució és desnormalitzar totes les dades i d'aquesta manera, en les consultes, s'ubiquen en una mateixa ranura. L'estalvi d'emmagatzematge no és un problema per un sistema que pot escalar a PetaByte. Tot i això, pot suposar un problema si es vol agrupar o ordenar per un camp quan hi ha una relació d'1 a molts.

Per optimitzar la desnormalització de les dades, Google proposa crear camps jerarquitzats (nested fields) o repetits. Aquests camps són de tipus registre i permeten mantenir la relació sense l'impacte d'un esquema relacional. Evidentment aquesta solució només és útil si no hi ha molts canvis en l'esquema d'estrella o modificacions en els registres i en el nostre treball no es produirà cap dels dos casos.

Per la implementació hem desnormalitzat totes les dimensions però creant camps jerarquitzats per no perdre l'estructura de les zones i les segmentacions. Els camps numèrics i mesurables s'han afegit també com un camp jerarquitzat repetible. Per poder processar eficientment les mètriques agrupades per les diferents dimensions. L'esquema implementat a BigQuery és mostra a la següent figura:

Table Details: Campaigns

Schema | Details | Preview

_PARTITIONTIME	TIMESTAMP	NULLABLE	This pseudo column contains a timestamp for the start of the day (in UTC) in which the data was loaded. For the YYYYMMDD partition, this pseudo column will contain the value TIMESTAMP('YYYY-MM-DD').
Platform	STRING	NULLABLE	Xarxes Socials
Products	RECORD	NULLABLE	Describe this field...
Products.**Product**	STRING	NULLABLE	Producte de l'anunci
Products.**Family**	STRING	NULLABLE	Família del producte
Ages	RECORD	NULLABLE	Describe this field...
Ages.**Age**	STRING	NULLABLE	Rang d'edats
Genders	RECORD	NULLABLE	Describe this field...
Genders.**Gender**	STRING	NULLABLE	Sexe
Interests	RECORD	NULLABLE	Describe this field...
Interests.**Likes**	STRING	NULLABLE	Gustos
Zones	RECORD	NULLABLE	Describe this field...
Zones.**Zone**	STRING	NULLABLE	Zona geogràfica
Zones.**City**	STRING	NULLABLE	Ciutat
Zones.**ZipCode**	STRING	NULLABLE	Codi Postal
Ads	RECORD	REPEATED	Describe this field...
Ads.**Date**	DATETIME	NULLABLE	Data de l'anunci
Ads.**Prints**	INTEGER	NULLABLE	Nombre d'impressions
Ads.**Hits**	INTEGER	NULLABLE	Nombre de Clicks que s'han fet
Ads.**CTR**	FLOAT	NULLABLE	Proporció de Clicks respecte el nombre d'impressions

Figura 6: Captura amb els camps de la base de dades

4. Processos ETLs

Els processos d'extracció, transformació i càrrega (Extract, Transform and Load, ETL) ens ha de permetre omplir de dades BigQuery amb el model i esquema implementat. Cada font d'informació té unes característiques diferents que compliquen la connexió i adquisició de les campanyes a estudiar. No disposem d'una eina pròpia per fer la integració amb les xarxes socials però sí que tenim diferents productes que, segons l'enfocament que es vol aplicar, es poden fer servir per a aquests processos.

Mostrarem dos models de solucions: un de teòric i un altre que, pels recursos disponibles, és el que s'ha implementat. En el primer, l'idea, és capturar les dades directament de les xarxes socials mitjançant tasques programades que cridaran les funcions de canalització d'informació. Aquesta informació s'adaptarà i s'enviarà a BigQuery. En el segon model les dades primer són emmagatzemades en el núvol de Google (Cloud Storage) per a després, de forma manual, tractar-les i adaptar-les per enviar-les al nostre magatzem.

4.1 Model teòric

El model teòric pretén ser la base per un sistema amb solucions ETL que automatitzin les tasques d'extreure les dades, fer les transformacions i les carregues a les taules de Google BigQuery. Aquesta solució té l'estructura següent:

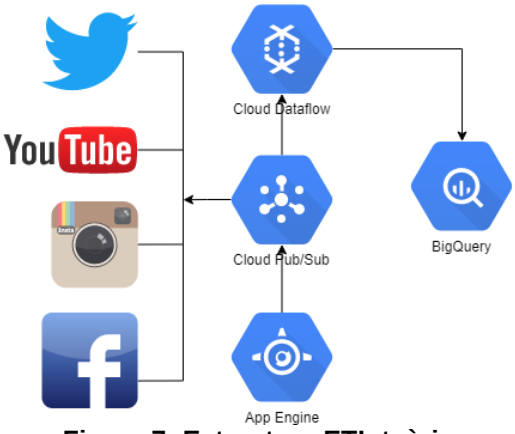


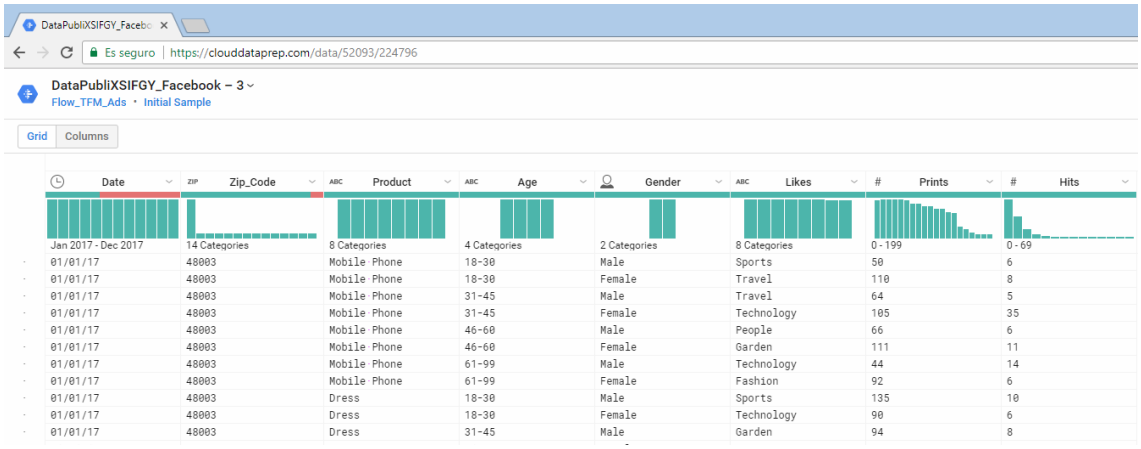
Figura 7: Estructura ETL teòric

- App Engine programa l'extracció de dades (amb tasques CRON) perquè es faci temporalment i així l'usuari no hagi de fer les actualitzacions manualment.
- Aquesta programació crida Cloud Pub/Sub per l'extracció de les dades i l'envio a Cloud Dataflow per processar-les. Es pot utilitzar Cloud Functions per realitzar les activacions de Pub/Sub.
- Dataflow fa les transformacions i la neteja de dades i les incorpora a Bigquery.

4.2 Solució implementada

La implementació feta dels processos ETLs treballa amb la hipòtesi de tenir les dades en un fitxer, en el nostre cas treballem amb un Excel de campanyes simulades i que es pujarà al núvol de Google per tractar-lo. Ens hem basat en la utilització de Cloud Dataprep que permet la preparació, transformació i neteja de dades d'una forma visual i àgil.

Dataprep agafa una mostra de les dades i les ensenya per pantalla separant-les en columnes i identificant els seus tipus. Mostra quantes i quines categories hi ha per columna i es marquen en vermell els valors que són invàlids pel tipus de dada. S'obté ràpidament una informació de les dades i proposa qualsevol tipus de transformació únicament polsant a sobre d'una categoria o columna.



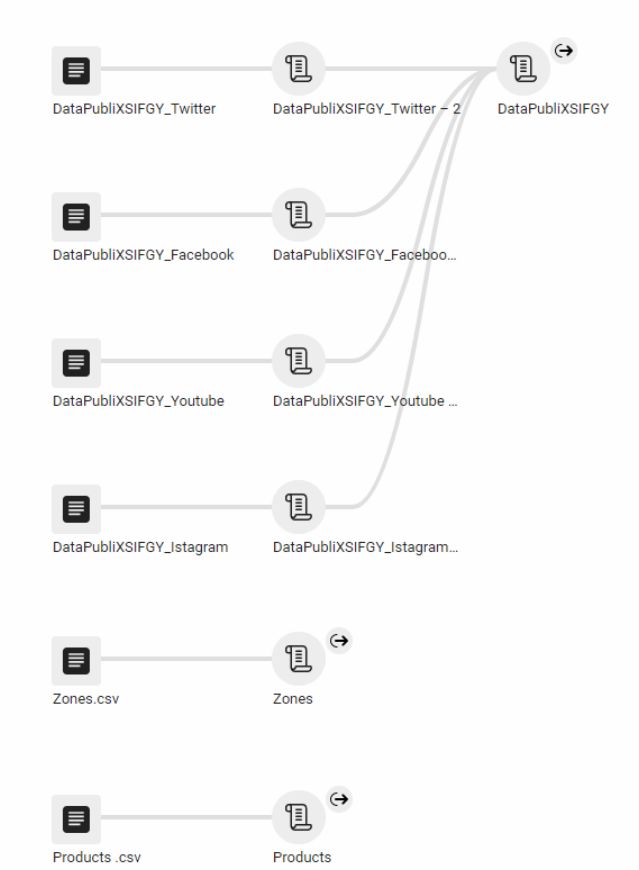
Aquesta eina, però, té la limitació que només tracta dades que estiguin allotjades a Google Cloud Storage (el magatzem de dades unificat) o en BigQuery. L'estructura de la solució implementada és una mica diferent de l'anterior:

Figura 8: Estructura ETL implementat

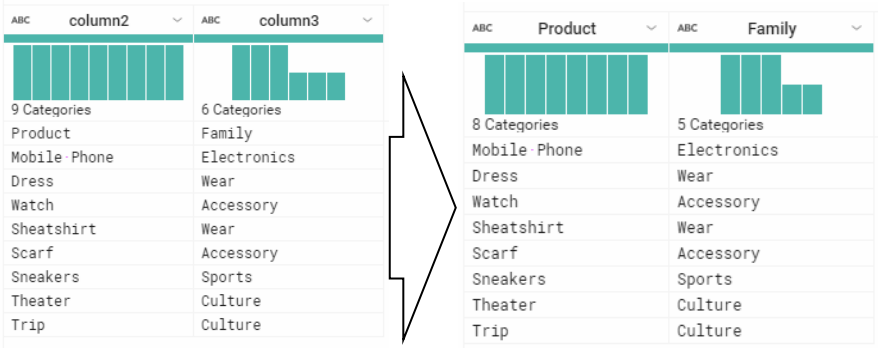
Aquests processos no són automàtics però, si s'entenen bé, són molt fàcils de seguir. Es divideixen en les següents passes:

1. **Preparació de l'Excel**: Dataprep pot carregar directament arxius de Microsoft Excel però pot donar problemes si tenen mides de més de 35MB. La solució que dóna és convertir les pestanyes en arxius CSV. Com l'Excel del treball pesava gairebé 58MB, es va crear un CSV per a cada pestanya.
2. **Creació dels fluxos de canvis**: es carreguen tots els arxius CSV i es crea un flux (Flow) on a cada set de dades se li aplicarà les corresponents transformacions o receptes (Recipe), com

s'anomena en Dataprep. Les dades de les campanyes publicitàries s'unificaran al final del flux. Sempre és millor fer les unions la final perquè així es té l'avantatge de processar els canvis en paral·lel.



3. **Transformació "Products"**: Dataprep no detecta la primera fila com a nom de columna. S'ha d'indicar, com un pas en la recepta, que agafi la primera fila com a capçalera.



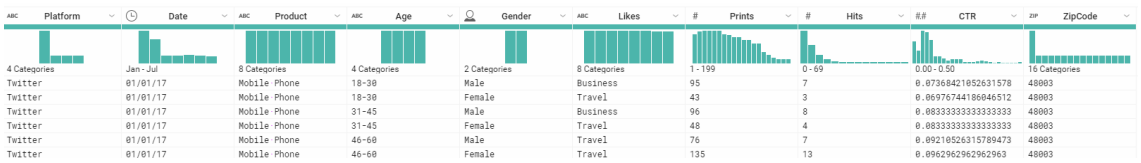
column2	column3
9 Categories	6 Categories
Product	Family
Mobile Phone	Electronics
Dress	Wear
Watch	Accessory
Sheatshirt	Wear
Scarf	Accessory
Sneakers	Sports
Theater	Culture
Trip	Culture

Product	Family
8 Categories	5 Categories
Mobile Phone	Electronics
Dress	Wear
Watch	Accessory
Sheatshirt	Wear
Scarf	Accessory
Sneakers	Sports
Theater	Culture
Trip	Culture

4. **Transformació "Zones"**: no s'ha de fer cap tipus de canvi.
5. **Transformacions en els fitxers de les dades de publicitats**: per cadascun es creen dues noves columnes una amb el resultat d'una fórmula (la fórmula és el nom de la xarxa social) i altra amb el càlcul del CTR (no ens ve donat i d'aquesta manera ens estalvien recalcular-lo a cada consulta). S'eliminen els valors de eficiència CTR que són nuls (provocats per anuncis sense cap impressió).

Es pot veure que hi ha codis postals que són incorrectes. Això és perquè al crear l'arxiu CSV, els codis que comencen per 0 han perdut el primer dígit. Es fa unes transformacions per recuperar-lo afegint un 0 a l'esquerra i després agafant els 5 dígits de la dreta.

6. **Unió de fitxers**: a l'últim pas o recepta que hi ha al flux es fa una unió de totes les taules d'anuncis. També es canvia les dates al format americà de mesos, dies i anys. Per fer-ho primer es canvia el tipus de dades a cadena i després a format data però indicant que els dies van primer. Al final no hi hauria d'haver cap dada incorrecta (totes les categories en verd).



7. **Creació de la sortida de l'execució**: Dataprep pot desar el resultat de les transformacions a BigQuery, però no pot generar camps jerarquitzats (o almenys jo no he trobat com fer-ho) i la sortida no és vàlida pel nostre model.
   El que es farà és enviar totes les dades a taules independents, per posteriorment fer la transformació al nostre model mitjançant una

consulta de BigQuery. Es creen tres taules: Zones, Productes i DataPubliXSIFGY.

8. **Consulta per insertar al model**: es crea una consulta en SQL i el seu resultat és enviat directament al nostre model. La consulta i la configuració a BigQuery és la següent:

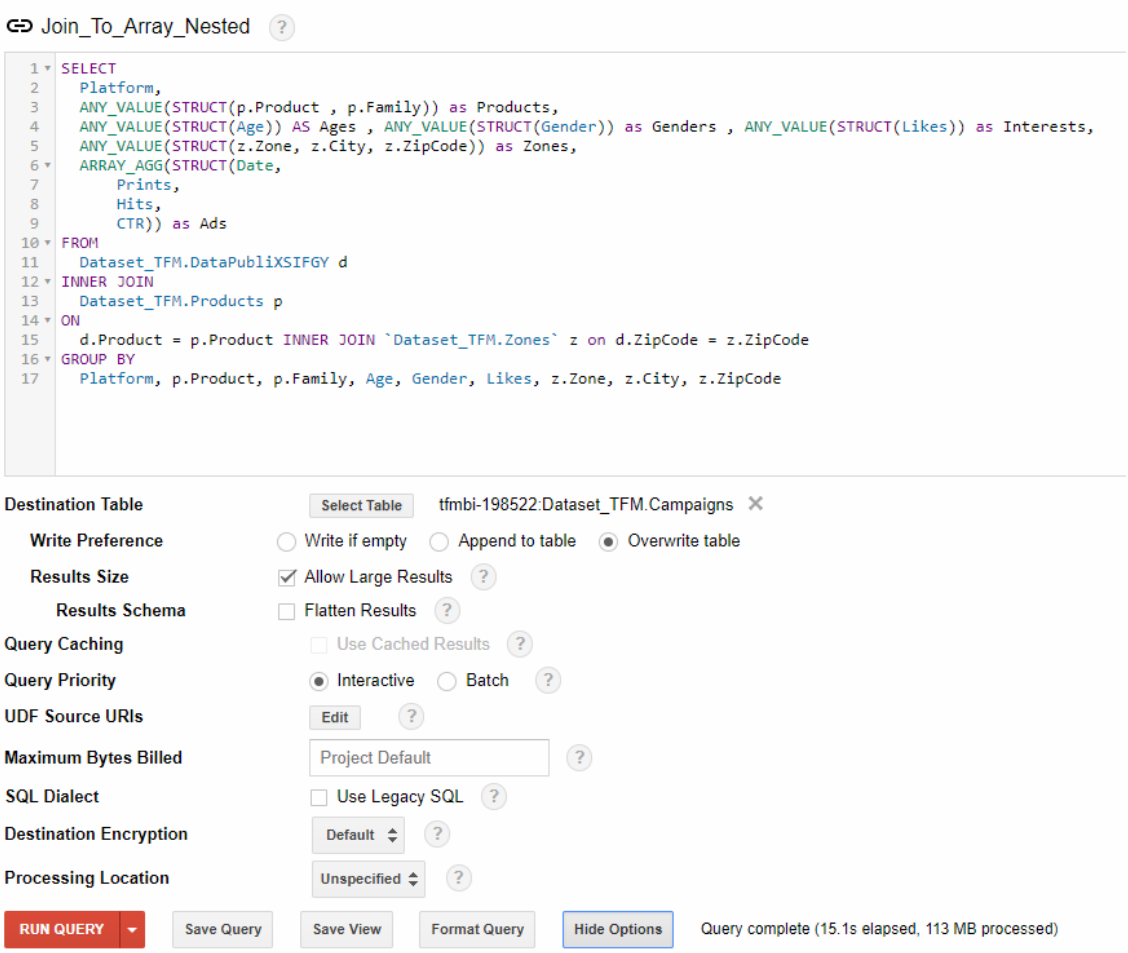


Figura 9: Captura amb la consulta per inserir les dades

Un cop s'han fet tots els passos ja disposem de les dades llistes per ser consultades i analitzades per les aplicacions BI.

5. Definició i optimització de consultes

Per a qualsevol solució d'intel·ligència de negoci tradicional, un cop ja es té totes les dades en el nostre dipòsit, el següent pas consisteix en la creació de cubs OLAP per augmentar el rendiment de les consultes a nivell d'agregació. En BigQuery aquesta solució no existeix, o almenys Google no dóna aquesta possibilitat. Aquest apartat vol explicar com és possible crear un sistema OLAP sense tenir cubs i perquè no és necessari.

5.1 Solucions OLAP i BigQuery

Els sistemes processament analític en línia o OLAP (On-Line Analytical Processing) són els que es fan servir en l'àrea de BI. Es caracteritzen per ser dissenyats per agilitzar consultes en grans quantitats de dades. Es poden identificar els següents tipus:

- **ROLAP o processament analític en línia relacional**: solució basada i construïda sobre una base de dades relacional. Tenen els avantatges d'una base de dades relacional, donant més escalabilitat que una solució MOLAP però que fa que sigui necessari la generació d'índexs, abans d'executar les consultes, perquè sigui ràpid. En molts casos s'han de compilar molts índexs per cobrir les necessitats de les consultes.
- **MOLAP o processament analític multidimensional en línia**: en aquesta solució les dimensions i els fets o valors agrupats s'emmagatzemen en matrius multidimensionals. Aquestes matrius són els anomenats cubs OLAP. Estan basats en les dimensions predefinides durant la fase de disseny. Després de crear-los, els usuaris i analistes poden obtenir ràpidament resultats. Els cubs disposen d'unes operacions bàsiques que milloren l'accessibilitat de la informació com són: "drill down" i "roll up" per moure la vista a un nivell major o menor de detall, "Slice" per tallar el cub en un subcub i així fer que l'usuari se centri només en una àrea o "Dice" per rotar el cub i poder veure les dades des d'una altra perspectiva o jerarquia. La debilitat dels cubs és que els enginyers de TI han d'invertir temps i recursos en el seu disseny i la seva construcció. Els analistes han d'esperar molt de temps quan demanen consultes noves o Ad Hoc i, si hi ha hagut algun canvi en l'esquema poden provocar errors en els dissenys implementats.
- **BigQuery**: Google descriu el seu producte com un sistema OLAP, alguns experts parlen d'una base de dades analítica, i està dissenyat per aplicacions de processament analític i BI que requereixen consultes complexes en conjunts de dades grans. Realment no és un sistema OLAP tradicional. La seva similitud radica en la seva capacitat per respondre ràpidament a consultes multidimensionals. No hi ha cubs i els seus esquemes no es basen en estrella o flocs de neu sinó en una taula aplanada gran i en realitzar consultes Ad Hoc.
  Internament, emmagatzema les dades en un format en columnes propietari anomenat Capacitor (Google, 2016), evolucionant en paral·lel amb el motor de consultes fent que s'aprofiti del coneixement de la

distribució de les dades per optimitzar seva l'execució. BigQuery utilitza patrons d'accés per determinar la quantitat òptima de fragments físics i com estan codificats.
Les dades s'emmagatzemen físicament en un sistema d'arxius distribuïts, també de propietat de Google, anonemat Colossus (Google, 2010), l'última versió de GFS. Assegura la durabilitat replicant les dades en fragments redundants de discos físics i amb l'ús del mètode de codi d'esborrat.

BigQuery pot obtenir millors resultats que els que obtindríem en un magatzem de dades tradicional. Òbviament és una suposició perquè s'hauria de fer diversos test de rendiment amb altres sistemes.

L'ús de cubs en la part superior d'un magatzem pot suposar que, canviar el magatzem, signifiqui un efecte dominó de canvis en tota la solució. I més important que el canvi és la velocitat real en què es poden realitzar aquests canvis. Amb un enfocament de lliurament de les dades "àgil", com es va explicar anteriorment per un sistema modern, no es pot reaccionar prou ràpid als requisits de l'entorn empresarial si s'ha de redissenyar i recalcular els cubs.

Tot i que aplicar una capa OLAP per sobre de BigQuery probablement complicaria les coses i pondria límits a l'anàlisis de dades, s'ha volgut fer un petit estudi per veure les possibilitats que trobem. Els sistemes ROLAP o MOLAP poden ser que no siguin adequats per consultes Ad Hoc o en anàlisis de prova i error, però és possible que es vulgui fer servir un servidor específic OLAP, on es precomputen les diferents agregacions i s'emmagatzemen les dades organitzades i indexades que no un motor on totes les agregacions de les consultes s'han de fer sobre la marxa.

5.2 Optimització en els accessos a BigQuery

Google ofereix recomanacions per optimitzar les consultes i els accessos a BigQuery. Es vol obtenir una eficiència en el temps d'execució però també en la quantitat de dades que es mouen en la construcció dels resultats. Els quadres de comandament i els informes es generaran amb Data Studio. Tot i que es pot realitzar una connexió directa amb BigQuery i Data Studio emmagatzema els resultats en la memòria cau, hi ha situacions on interessa reduir la quantitat d'informació que s'enviarà.

Cada mes es tarifa la memòria que s'ha utilitzat en les consultes, el preu és molt baix (5$ per cada TB i el primer és gratuït) i la quantitat d'informació que tenim en el treball no és molt alta però es vol garantir uns accessos eficients per un projecte a futur o, com a mínim, assentar unes bases de com millorar les consultes.

L'optimització que s'ha utilitzat, part de la idea que els informes mostrats són periòdics, típic en sistemes on els usuaris de negoci accedeixen als resultats mensuals. Això vol dir que no s'agafarà les dades en temps real ni les més actuals, sinó que periòdicament es generarà unes vistes amb la informació necessària per generar els informes. Pel nostre treball de màster aquesta

optimització potser no té gaire sentit, es vol fer un estudi puntual sobre l'efectivitat dels anuncis en diferents jerarquies, però ens sembla interessant realitzar-la per tal d'aconseguir un model de BI adaptable a qualsevol necessitat.

Les taules es poden dividir en particions diàries i així es pot filtrar la informació que es vol agrupar periòdicament d'una manera més ràpida. Per actualitzar les vistes ens ajudarem amb Google App Engine. El procés tindrà la següent estructura:

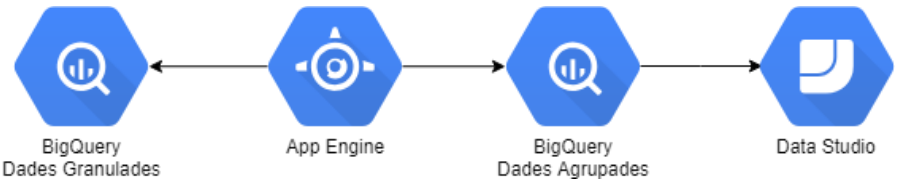


Figura 10: Estructura optimització per a Dashboards

La primera pregunta analítica que es vol resoldre amb el treball de màster és l'efectivitat de les regions o ciutats i si tenen alguna relació amb els productes o famílies. La farem servir per explicar els passos de la millora.

1. Es crea la taula de report amb la informació agrupada a data actual en un nou conjunt de dades.

Table Details: Effectiveness_Zones_Products

Schema | Details | Preview

Zone	STRING	NULLABLE	Describe this field...
City	STRING	NULLABLE	Describe this field...
Performance	FLOAT	NULLABLE	Describe this field...
Product	STRING	NULLABLE	Describe this field...
family	STRING	NULLABLE	Describe this field...

Figura 11: Camps del report

2. S'insereixen les dades actuals a la taula nova mitjançant la següent consulta

```
SELECT Zones.Zone , Zones.City , AVG(Ad.CTR) as Performance,Products.Product ,
Products.Family
FROM  `Dataset_TFM.Campaigns` c CROSS JOIN UNNEST(c.Ads) AS Ad
GROUP BY Zones.Zone , Zones.City , Products.Product , Products.Family
```

3. Es crea la vista que seleccionarà les dades inserides durant 7 dies (farem una periodicitat setmanal) per enviar-les a la taula de reports. S'anomenarà "Performance_Zones_Products_Weekly".

```
SELECT Zones.Zone , Zones.City , AVG(Ad.CTR) as Eficiencia,Products.Product ,
Products.Family
FROM  `treball-final-master.Dataset_TFM.Campaigns` c CROSS JOIN UNNEST(c.Ads) AS Ad
WHERE _PARTITIONTIME >= TIMESTAMP_SUB(CURRENT_TIMESTAMP(), INTERVAL 7 DAY)
   AND _PARTITIONTIME < TIMESTAMP_TRUNC(CURRENT_TIMESTAMP(), DAY)
GROUP BY Zones.Zone , Zones.City , Products.Product , Products.Family
```

4. Es crea una petita aplicació, feta amb Python, que llegeixi les dades actuals i les enviï agrupades a la nova taula periòdicament. Bàsicament cridem la vista creada per inserir el seu resultat a la taula de report. Una part del codi font és la següent:

```
def Async_Zones_Products():
    # [START build_service]
    # Construct the service object for interacting with the BigQuery API.
    bigquery_client = bigquery.Client()
    # [END build_service]

    # Submit the job and wait for it to complete.
    query = ('SELECT * FROM tfmbi-198522:Dataset_Reports.Performance_Zones_Products_Weekly')
    query_job = async_query(
        bigquery = bigquery_client,
        project_id = 'tfmbi-198522',
        query = query,
        datasetId = 'Dataset_Reports',
        dataset_projectId = 'tfmbi-198522',
        destination_tableId = 'Effectiveness_Zones_Products',
        batch=True,
        use_legacy_sql=False)

    poll_job(bigquery_client, query_job)
```

La funció envia una petició asíncrona d'una consulta creada amb el codi següent:

```
def async_query(
        bigquery, project_id, query,datasetId,dataset_projectId,destination_tableId,
        batch=False, use_legacy_sql=False):
    # Generate a unique job ID so retries
    # don't accidentally duplicate query
    job_data = {
        'jobReference': {
            'projectId': project_id,
            'jobId': str(uuid.uuid4())
        },
        'configuration': {
            'query': {
                'query': query,
                'priority': 'BATCH' if batch else 'INTERACTIVE',
                'useLegacySql': use_legacy_sql,
                'writeDisposition': 'WRITE_APPEND',
                    'destinationTable': {
                        'datasetId': datasetId,
                        'projectId': dataset_projectId,
```

```
                    'tableId': destination_tableId
                }
            }
        }
    }
    return bigquery.jobs().insert(
        projectId=project_id,
        body=job_data).execute()
```

5. I es crea una tasca CRON per executar l'aplicació cada 7 dies. L'arxiu "cron.yaml" és el següent:

```
cron:
- description: Importa les dades agrupades pels informes de efectivitat en productes i zones
  url: /events/zones_products_import
  schedule: every 168 hours
  timezone: Europe/Madrid
```

Ara les aplicacions BI es poden connectar a les taules de reports estalviant-nos processar tots els registres per mostrar informes que, sovint, són consultats els mateixos per diverses persones. També ens dóna la possibilitat de definir permisos d'accés a consultes concretes.

5.3 Proves model

Per finalitzar, s'ha volgut fer unes proves de rendiment en els diferents models de dades plantejats i amb les taules de reports periòdics per veure el temps d'execució i la quantitat d'informació manegada.

La consulta per fer les proves és l'efectivitat mitjana de les ciutats de la zona costanera agrupada per families de productes.

- Amb un model normalitzat, el resultat d'una consulta amb unions de taules és de 5,3 segons i 33,8 MB processats.

```
SELECT z.City, p.Family, AVG(d.CTR) As Efectivitat FROM [tfmbi-
198522:Dataset_TFM.DataPubliXSIFGY] d
INNER JOIN [tfmbi-198522:Dataset_TFM.Products] p ON d.Product = p.Product
INNER JOIN [tfmbi-198522:Dataset_TFM.Zones] z ON d.ZipCode = z.ZipCode
WHERE z.Zone = 'Coast'
GROUP BY z.City, p.Family
```

- Amb un model desnormalizat i amb camps jerarquitzats el temps d'execució i la quantitat de dades processades és de 1,9 segons i 12,1 MB.

```
SELECT Zones.City , Products.Family , AVG(Ad.CTR) as Performance
FROM  `Dataset_TFM.Campaigns` c CROSS JOIN UNNEST(c.Ads) AS Ad
```

```
WHERE Zones.Zone = 'Coast'
GROUP BY Zones.City , Products.Family
```

- Amb una taula de dades agrupades per a reports el resultat és de 2 segons i 5.35 KB.

```
SELECT City , Family , AVG(Performance) as Efectivitat FROM `tfmbi-
198522.Dataset_Reports.Effectiveness_Zones_Products`
WHERE Zone = 'Coast'
GROUP BY City , Family
```

Es pot apreciar que amb un model desnormalitzar i amb camps lligats el temps d'execució gairebé és el mateix que amb la taula preparada per informes, o com a mínim per la nostra quantitat de registres. Demostra que l'optimització de llistats periòdics no suposa una eficiència molt gran per després muntar les nostres aplicacions BI.

6. Generació d'informes i analítica.

L'objectiu últim del treball final de màster és resoldre les qüestions analítiques. La millor manera per veure la informació o poder extreure el coneixement necessari per resoldre-les és de forma gràfica. Per això s'ha triat una eina com Data Studio perquè permet la generació de taulers i quadres de comandament (Dashboard) gràfics.

La mètrica que serà mostrada en els diferents "Dashboards" serà la mitjana dels CTRs, que és el millor indicador per representar l'efectivitat dels anuncis. S'intentarà que sigui visualment llegible i que permeti que l'usuari pugui comparar fàcilment la informació.

Tots els gràfics i informes es poden consultar a https://datastudio.google.com/open/1GUrx29R7LDqUFbBaWF3obd_0Y5A3M6yn

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

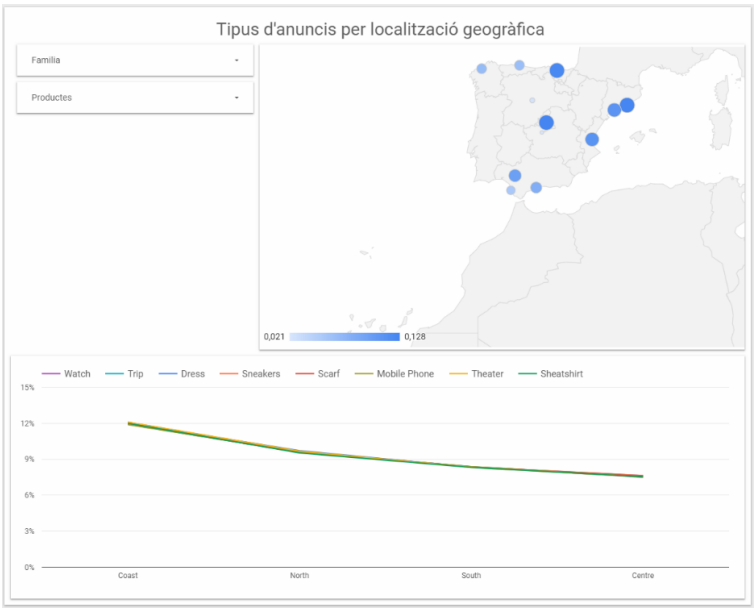


Figura 12: Captura Dashboard 1

Per poder extreure la relació entre els tipus de productes i la seva localització geogràfica s'ha decidit col·locar les dades dins d'un mapa juntament amb un gràfic detallat per productes. Així es pot veure fàcilment que les ciutats amb millors indicadors són Madrid, Barcelona i Bilbao, seguides per Tarragona i València.

Si ens fixem en el gràfic, les línies de cada producte són gairebé idèntiques per a cada regió. Això ens fa entendre que no hi ha cap producte o família que sigui millor per una zona en concret, sinó que l'efectivitat ve donada per la localització de l'anunci.

S'ha afegit uns controls de filtres per poder seleccionar productes i consultat si un tipus concret té més efectivitat en una ciutat concreta. Les diferències que trobem no són molt significatives.

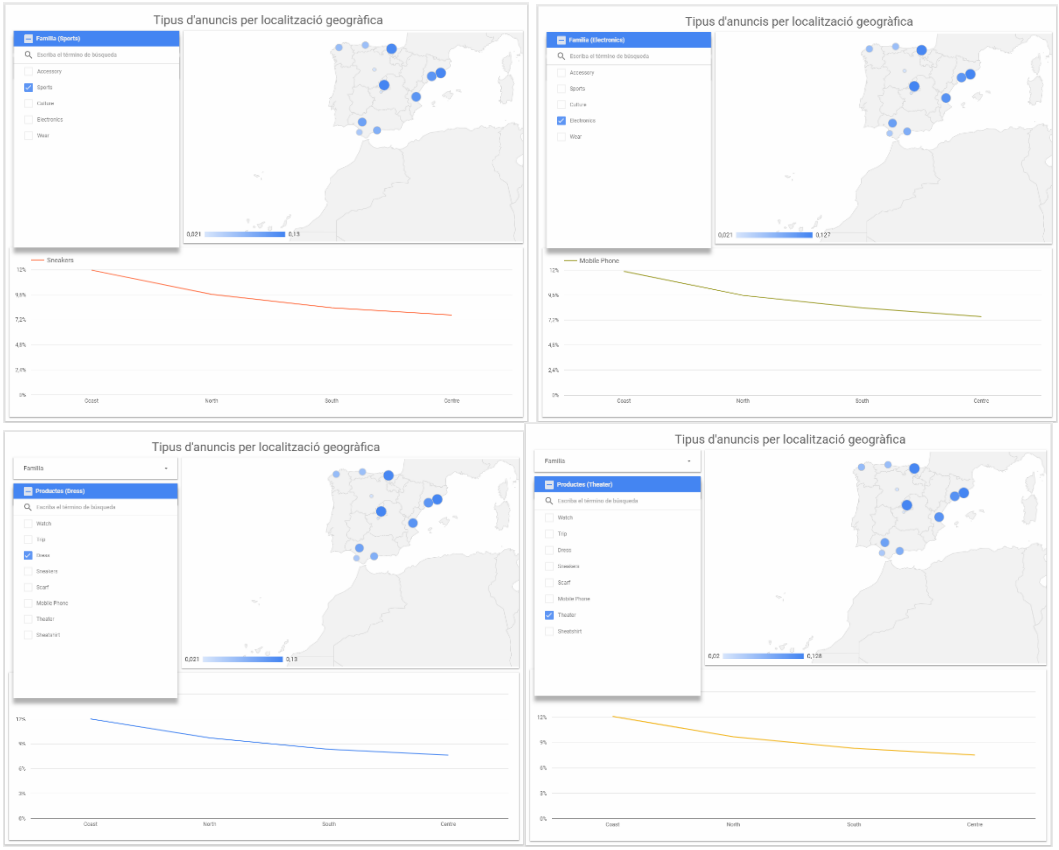


Figura 13: Captures Dashboard 1 amb diversos filtres

6.2. Existeixen una relació entre la millora dels indicadors d'efectivitat amb algun segment de la població objectiu?

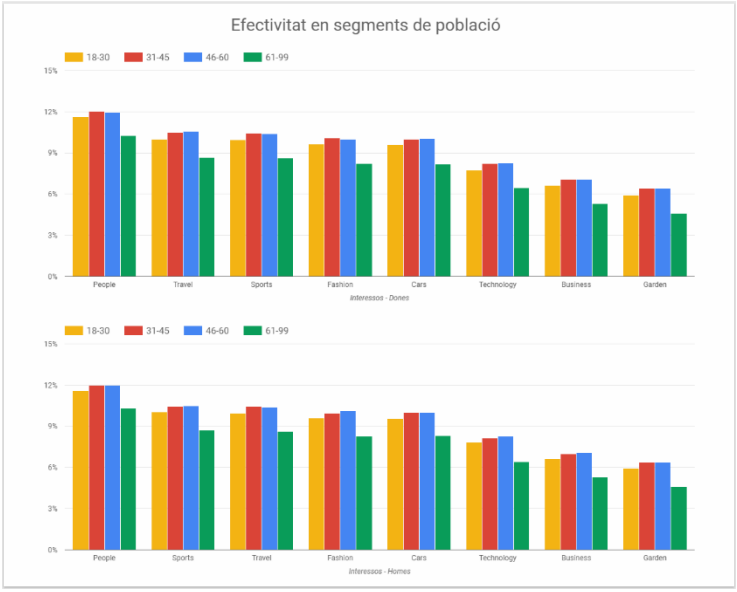


Figura 14: Captura Dashboard 2

La millor forma que hem cregut per mostrar l'efectivitat en els segments de població ha sigut un gràfic de barres. Per limitacions de Data Studio se n'han creat dos, un pel segment de dones i altre pels homes.

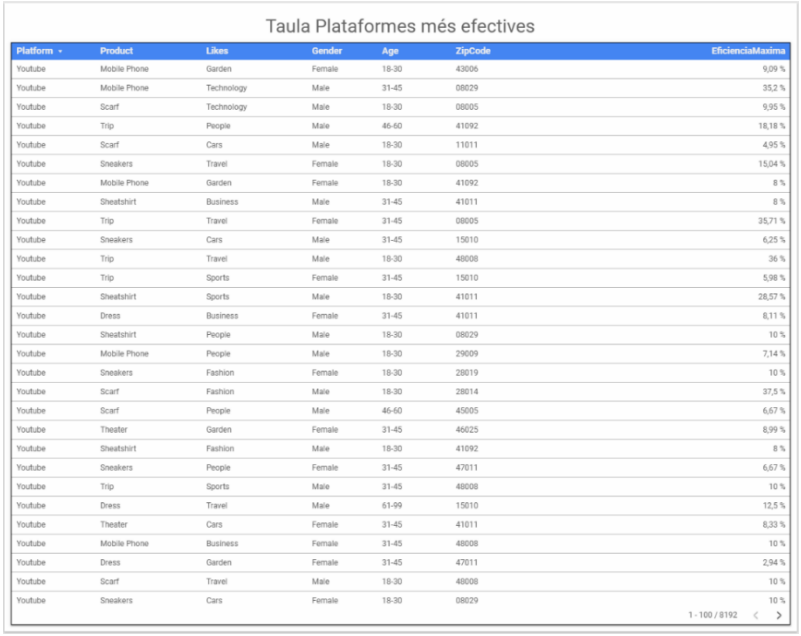
No s'aprecia una diferència clara entre sexes, només una petita millora en gustos esportius pels homes, però sí que podem dir que les franges d'edat d'entre 31 i 60 tenen més possibilitats que facin "Clicks" als anuncis i que la població amb gustos "People", "Travel" i "Sports" són més actius en les campanyes que els segments de "Garden", "Business" o "Technology".

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

Per aquesta qüestió podríem haver fet una pantalla semblant a la del punt següent, on l'usuari pogués filtrar per les condicions desitjades i així veure quina plataforma és millor, però ens ha semblat més interessant realitzar una consulta Ad-Hoc que ens retorni, per a cada conjunt de condicions, quina és la millor plataforma per mostrar un anunci. La consulta és la següent:

```
WITH tefimax as (
SELECT Platform , Products.Product, Ages.Age , Genders.Gender , Interests.Likes ,
Zones.ZipCode , AVG(Ad.CTR) as  EficienciaMaxima,
ROW_NUMBER() OVER (PARTITION BY Products.Product, Ages.Age , Genders.Gender , Interests.Likes
, Zones.ZipCode ORDER BY AVG(Ad.CTR) DESC) row_num
FROM  `tfmbi-198522.Dataset_TFM.Campaigns` c CROSS JOIN UNNEST(c.Ads) AS Ad
GROUP BY Platform, Products.Product, Ages.Age, Genders.Gender, Interests.Likes, Zones.ZipCode
)
SELECT Platform, Product, Age, Gender, Likes, ZipCode, EficienciaMaxima FROM tefimax
where row_num = 1
```

Des de Data Studio es mostra el resultat mitjançant una taula.



Taula Plataformes més efectives

Platform	Product	Likes	Gender	Age	ZipCode	EficienciaMaxima
Youtube	Mobile Phone	Garden	Female	18-30	43006	9,09 %
Youtube	Mobile Phone	Technology	Male	31-45	08029	35,2 %
Youtube	Scarf	Technology	Male	18-30	08005	9,95 %
Youtube	Trip	People	Male	46-60	41092	18,18 %
Youtube	Scarf	Cars	Male	18-30	11011	4,95 %
Youtube	Sneakers	Travel	Female	18-30	08005	15,04 %
Youtube	Mobile Phone	Garden	Female	18-30	41092	8 %
Youtube	Sheatshirt	Business	Male	31-45	41011	8 %
Youtube	Trip	Travel	Female	31-45	08005	35,71 %
Youtube	Sneakers	Cars	Male	31-45	15010	6,25 %
Youtube	Trip	Travel	Male	18-30	48008	36 %
Youtube	Trip	Sports	Female	31-45	15010	5,98 %
Youtube	Sheatshirt	Sports	Male	18-30	41011	28,57 %
Youtube	Dress	Business	Female	31-45	41011	8,11 %
Youtube	Sheatshirt	People	Male	18-30	08029	10 %
Youtube	Mobile Phone	People	Male	18-30	29009	7,14 %
Youtube	Sneakers	Fashion	Female	18-30	28019	10 %
Youtube	Scarf	Fashion	Male	18-30	28014	37,5 %
Youtube	Scarf	People	Male	46-60	45005	6,67 %
Youtube	Theater	Garden	Female	31-45	46025	8,99 %
Youtube	Sheatshirt	Fashion	Male	18-30	41092	8 %
Youtube	Sneakers	People	Female	31-45	47011	6,67 %
Youtube	Trip	Sports	Male	31-45	48008	10 %
Youtube	Dress	Travel	Male	61-99	15010	12,5 %
Youtube	Theater	Cars	Female	31-45	41011	8,33 %
Youtube	Mobile Phone	Business	Female	31-45	48008	10 %
Youtube	Dress	Garden	Female	31-45	47011	2,94 %
Youtube	Scarf	Travel	Male	18-30	48008	10 %
Youtube	Sneakers	Cars	Female	18-30	08029	10 %

1 - 100 / 8192

Figura 15: Captura Dashboard 3

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

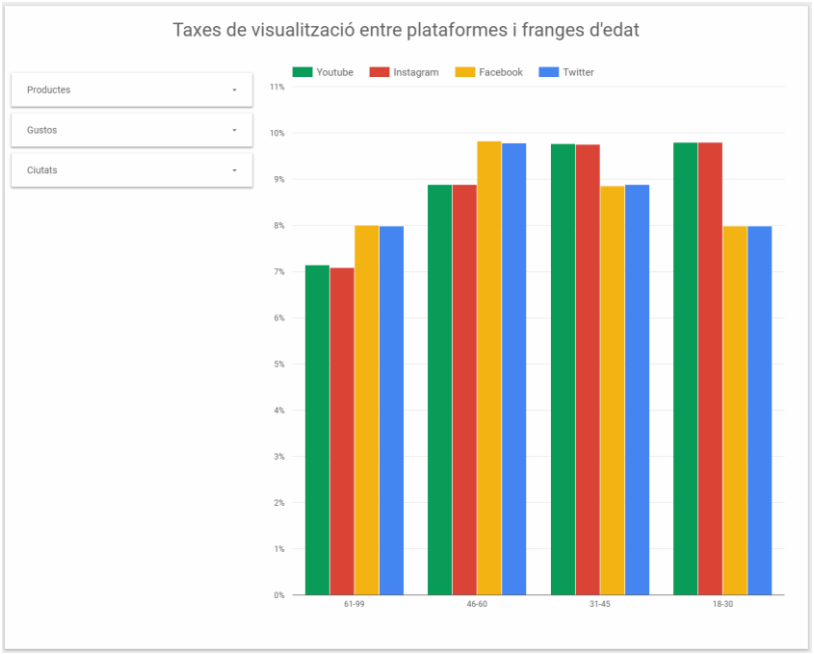


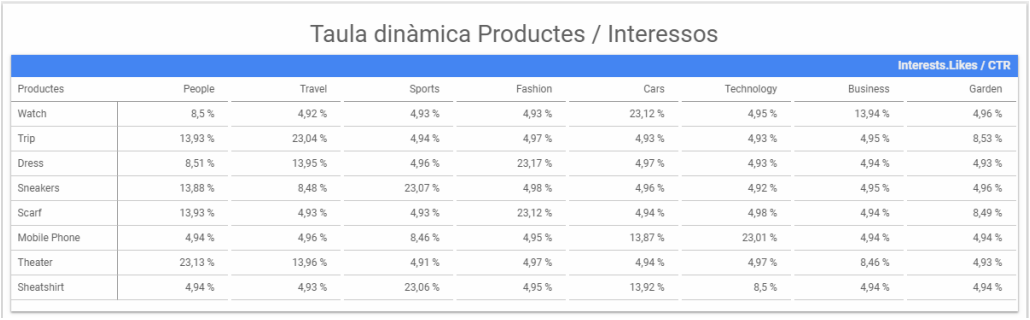
Figura 16: Captura Dashboard 4

Les plataformes Youtube i Instagram tenen millors taxes de visualització entre les franges 18-30 i 31-45 anys. Després sembla ser els usuaris es decanten per Facebook o Twitter. S'ha afegit una sèrie de controls per poder filtrar les dades i veure si n'hi ha situacions on aquesta relació no es compleix o, almenys, no és tan clara.

Figura 17: Captures Dashboard 4 amb diversos filtres

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

Per l'última pregunta s'ha utilitzat una taula dinàmica que ens relacioni les dimensions dels interessos dels usuaris amb els productes.



	Interests.Likes / CTR							
Productes	People	Travel	Sports	Fashion	Cars	Technology	Business	Garden
Watch	8,5 %	4,92 %	4,93 %	4,93 %	23,12 %	4,95 %	13,94 %	4,96 %
Trip	13,93 %	23,04 %	4,94 %	4,97 %	4,93 %	4,93 %	4,95 %	8,53 %
Dress	8,51 %	13,95 %	4,96 %	23,17 %	4,97 %	4,93 %	4,94 %	4,93 %
Sneakers	13,88 %	8,48 %	23,07 %	4,98 %	4,96 %	4,92 %	4,95 %	4,96 %
Scarf	13,93 %	4,93 %	4,93 %	23,12 %	4,94 %	4,98 %	4,94 %	8,49 %
Mobile Phone	4,94 %	4,96 %	8,46 %	4,95 %	13,87 %	23,01 %	4,94 %	4,94 %
Theater	23,13 %	13,96 %	4,91 %	4,97 %	4,94 %	4,97 %	8,46 %	4,93 %
Sheatshirt	4,94 %	4,93 %	23,06 %	4,95 %	13,92 %	8,5 %	4,94 %	4,94 %

Figura 18: Captura Dashboard 5

Es veu clarament que el coneixement del grup d'interès ajuda a aconseguir més "Clicks" com, per exemple, anuncis de viatges tenen més efectivitat si es mostren per usuaris amb gustos per viatjar (Travel) o si vols vendre un vestit no és bona idea personalitzar campanyes a amants de la tecnologia.

6.6. Altres qüestions analítiques

Amb les dades disponibles es poden analitzar més aspectes que els proposats per treball de màster. Com disposem de les dates de les campanyes es pot fer una anàlisi de l'evolució per veure si n'hi ha períodes on són més fluixes o si s'ha millorat en la qualitat de l'anunci (per exemples).

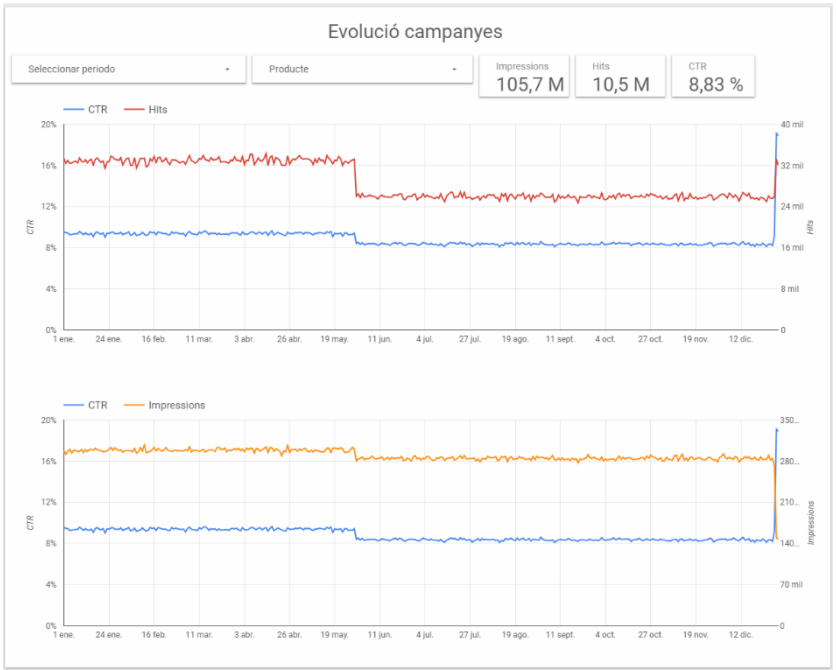


Figura 19: Captura Dashboard 6

Els canvis remarcables són una baixada de les impressions al juny, juntament amb una baixada amb l'efectivitat indicada pels CTRs i augment considerable dels dies 30 i 31 de desembre.

7. Conclusions

Quan vaig haver de triar una àrea on desenvolupar el treball final de màster ho vaig fer pensant que m'havia de donar nous coneixements, a part de la utilització dels ja apresos, i que fossin útils pel meu món laboral (a curt o llarg termini). Després de fer el treball puc dir que les meves expectatives s'ha complit. M'ha ajudat a entendre una àrea desconeguda fins ara i que em pot ser molt útil per desenvolupar projectes dins de la meva empresa. El producte està enfocat a l'anàlisi de campanyes publicitàries però es pot adaptar perfectament a qualsevol tipus d'estratègia.

El desconeixement de la matèria em va suposar dedicar-hi més temps a l'estudi i recerca dels conceptes de la intel·ligència de negoci, a vegades endarrerint el desenvolupament del treball, però sense suposar cap problema a l'hora de seguir la planificació inicial. L'única diferència en la metodologia va ser que vaig haver d'adaptar l'execució del projecte, tal com l'havia plantejat inicialment, per una forma una mica més iterativa en comptes d'una tradicional. Mirava d'avançar una etapa per veure possibles problemes i poder adaptar amb temps l'evolució del treball.

També es va veure condicionada pel desconeixement de les eines utilitzades. Sabia el funcionament de la plataforma de Google però no havia treballat mai amb aquests productes i vaig patir petits problemes, a vegades fent-me dubtar de la viabilitat del treball. Com el funcionament de Dataprep, en alguns casos no és tribal saber quines transformacions realitzar per tal de tenir totes les dades vàlides i s'ha de mirar totes les possibilitats o com es fa l'exportació directament a Bigquery amb un esquema de camps jerarquitzats (no he trobat la solució).

La principal errada en la planificació de les tasques va ser en la idea inicial de treballar amb cubs OLAP, per posteriorment trobar-me que amb Bigquery no existeix aquesta tecnologia. En els sistemes de Business Intelligence és un recurs molt característic per agilitzar les consultes i vaig creure que en el Data Warehouse proposat per Google també es podrien crear. Això abans de veure la definició dels sistemes moderns on, sembla ser, cada vegada tenen menys sentit la seva creació.

Per manca de temps i recursos no s'ha pogut fer un sistema completament operatiu. Falta la part d'extracció de les dades a fonts reals. Però m'ha ajudat a entendre com funciona i sé que es podria implementar en un futur. Estaria interessant afegir més mètriques, com per exemple el cost de les campanyes, per poder enriquir molt més el coneixement analític que dóna l'aplicació. Crec que crear un sistema d'intel·ligència de negoci amb un Data Warehouse amb només la plataforma de Google és factible. He trobat alguns components que no són tan madurs com altres solucions (com per exemple Data Studio que encara està en fase beta) però que poden encaixar perfectament en les necessitats de molts projectes.

8. Glossari

- Big Data: concepte que fa referència a un conjunt de dades tan gran que no es pot tractar amb les aplicacions tradicionals.
- Business Intelligence o BI: en català intel·ligència de negoci, conjunt de metodologies, aplicacions i tecnologies per extreure coneixement de la informació i donar valor a l'empresa.
- Consulta Ad-Hoc: consulta creada per un propòsit o fi específic.
- MapReduce: model de programació per processar grans quantitats de dades de forma paral·lela.
- On Premise: aplicacions o infraestructura que són a dins de l'empresa.

9. Bibliografia

Gartner Inc. Magic Quadrant for Analytics and Business Intelligence Platforms. (26 / Febrer / 2018). *Magic Quadrant for Analytics and Business Intelligence Platforms*. Recollit de https://www.gartner.com/document/3861464

Gartner Inc. Magic Quadrant for Data Management Solutions for Analytics. (Febrer / 2018). *Magic Quadrant for Data Management Solutions for Analytics*. Recollit de https://www.gartner.com/document/3855698

Gartner, Magic Quadrant for Analytics and Business Intelligence Platforms. (26 de Febrer de 2018). *Magic Quadrant for Analytics and Business Intelligence Platforms*. Obtenido de Magic Quadrant for Analytics and Business Intelligence Platforms: https://www.gartner.com/document/3861464

Gartner, Magic Quadrant for Data Integration Tools. (2017). *Magic Quadrant for Data Integration Tools*. Obtenido de Magic Quadrant for Data Integration Tools: https://www.gartner.com/document/3777464

Google. (29 de Juliol de 2010). *Storage Architecture and Challenges.* Obtenido de Storage Architecture and Challenges: https://cloud.google.com/files/storage_architecture_and_challenges.pdf

Google. (2016). *Inside Capacitor, BigQuery's next-generation columnar storage format.* Obtenido de Inside Capacitor, BigQuery's next-generation columnar storage format: https://cloud.google.com/blog/big-data/2016/04/inside-capacitor-bigquerys-next-generation-columnar-storage-format

Google, Inc. (2010). *Dremel: Interactive Analysis of Web-Scale Datasets.* Recollit de Dremel: Interactive Analysis of Web-Scale Datasets: https://static.googleusercontent.com/media/research.google.com/es//pubs/archive/36632.pdf

Kimball, R. (2013). *The Data Warehouse Toolkit: The Defi nitive Guide to Dimensional Modeling, Third Edition.* Wiley.

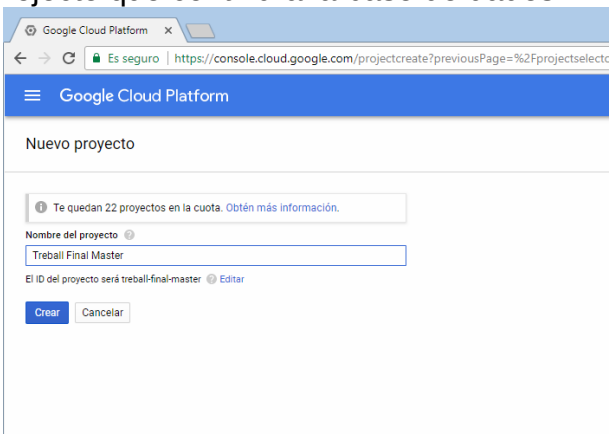
Sol, P. d. (24 de Maig de 2013). *admetricks blog*. Obtenido de admetricks blog: http://blog.admetricks.com/ctr-y-la-relevancia-en-el-marketing-online-que-es-para-que-se-usa-y-como-se-optimiza/

-

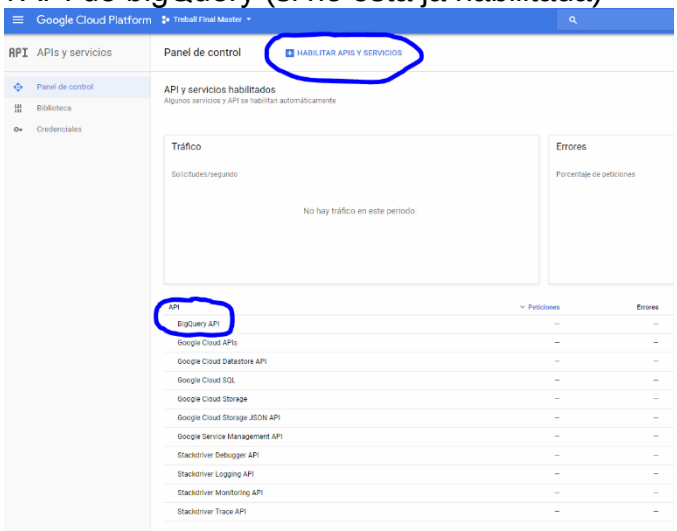
10. Annexos

A. Creació de una base de dades en BigQuery

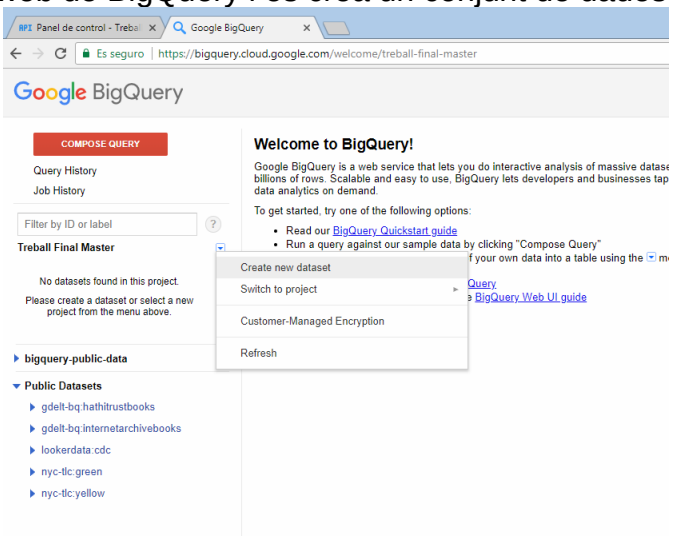
1. Es crea el projecte que contindrà la base de dades



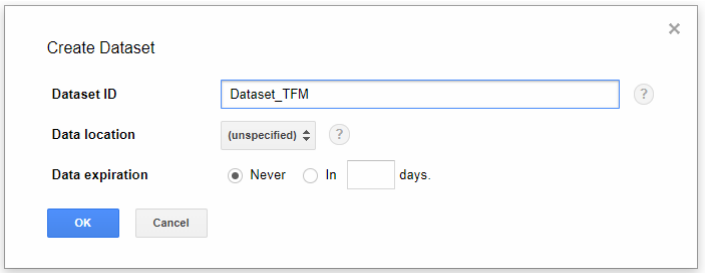
2. S'habilita l'API de bigQuery (si no esta ja habilitada)



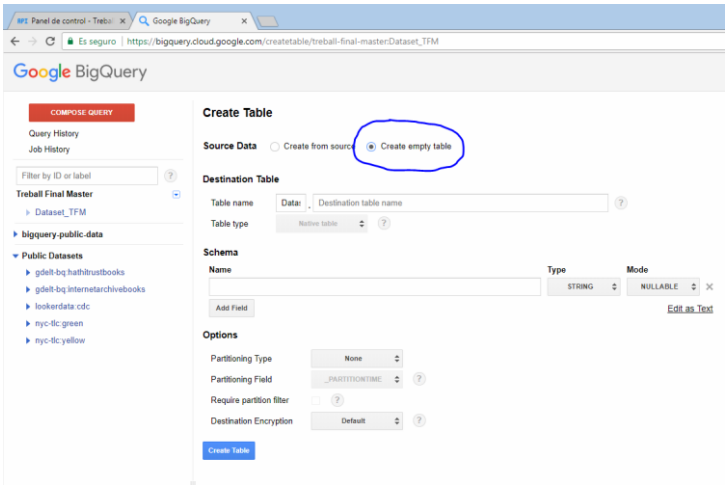
3. Obrim la web de BigQuery i es crea un conjunt de dades (Dataset)



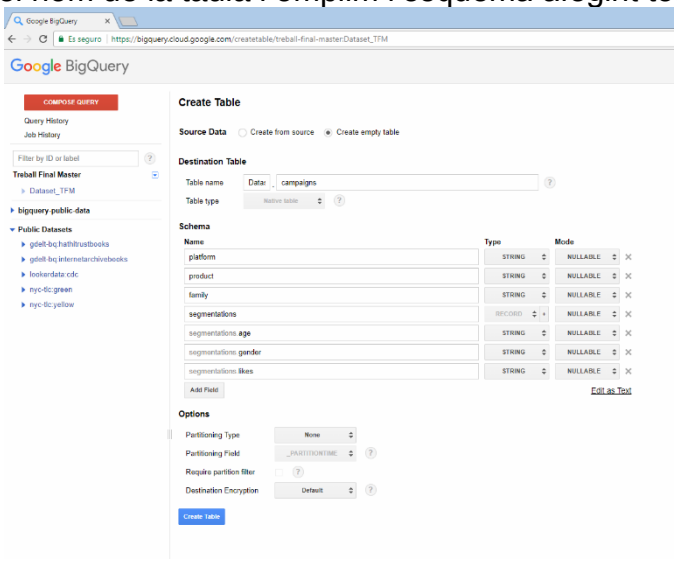
4. Podem triar on es volen desar les dades (US o EU), es recomanable que estiguin el més a prop possible de la aplicació que hi accedirà. I també podem posar una data de caducitat de la informació



5. Quan tenim el Dataset creat, amb el botó + que hi ha al costat es crea una taula. Es pot crear des d'un arxiu o una taula buida. Seleccionem "Crear taula buida".



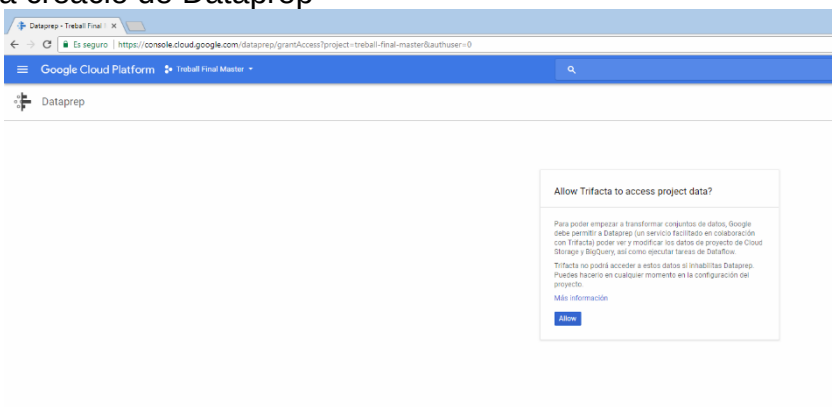
6. Escrivim el nom de la taula i omplim l'esquema afegint tots els camps



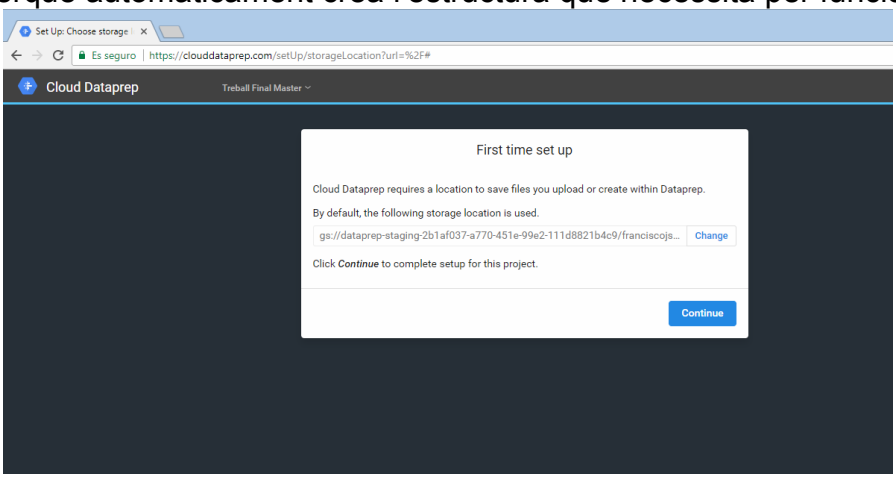
7. Per finalitzar s'ha de polsar el botó "Create Table".

B. Preparació i ús de Dataprep

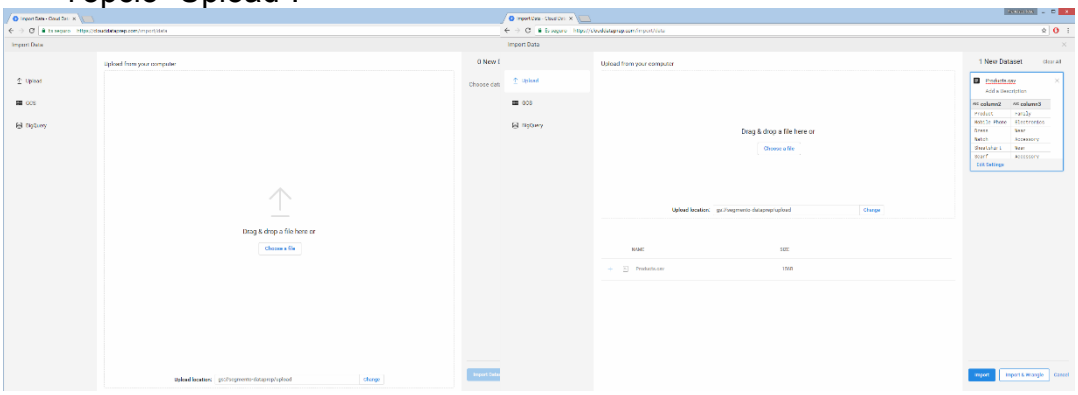
1. S'accedeix a Dataprep, si és la primera vegada en el projecte actiu s'ha de donar autorització a Trifacta (empresa sòcia i col·laboradora de Google en la creació de Dataprep



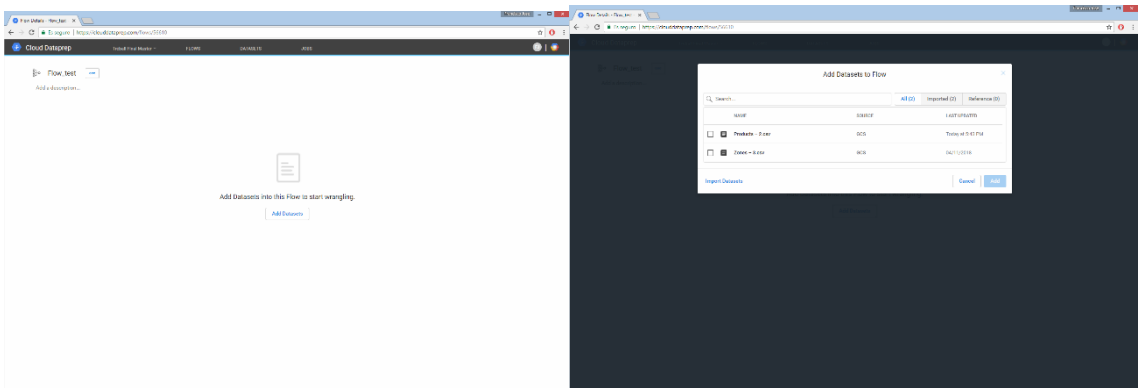
2. S'indica on s'emmagatzemen les dades que es pujaran per ser tractades dins de *Cloud Storage*. Es recomana deixar la localització per defecte perquè automàticament crea l'estructura que necessita per funcionar.



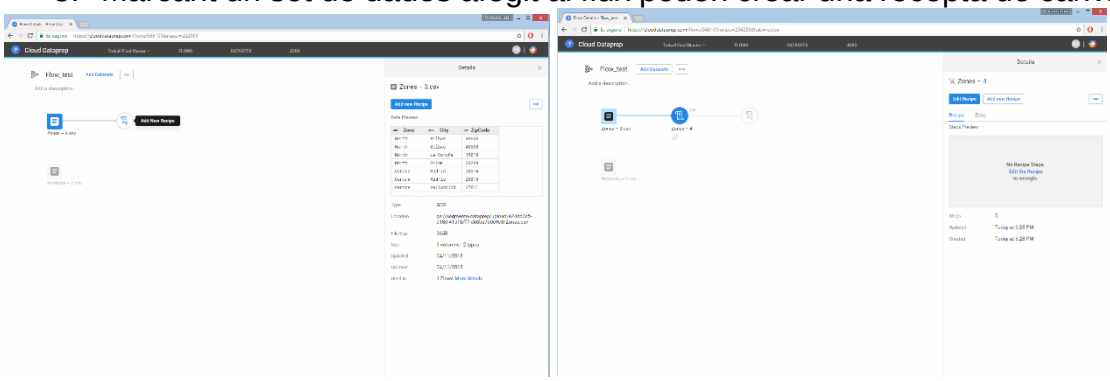
3. A la pestanya "DATASETS" s'importen les dades. Es poden agafar de Bigquery o d'arxius pujats al Cloud Storge o directament pujar-los amb l'opció "Upload".



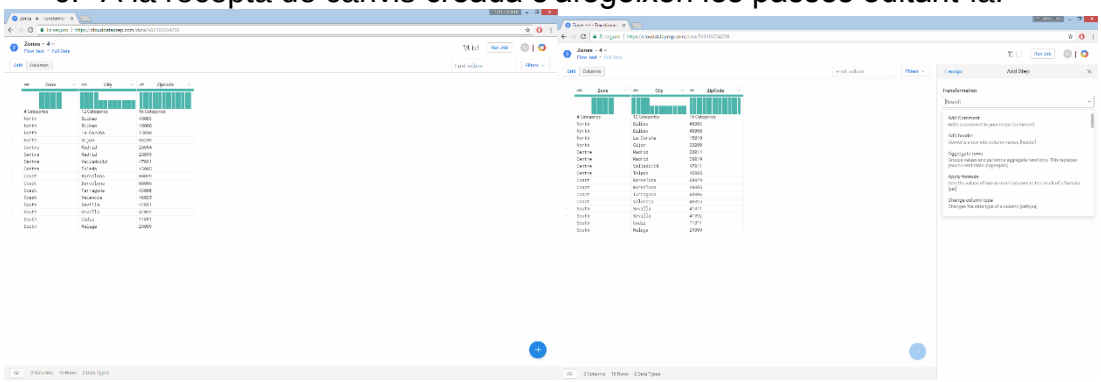
4. Quan es tenen totes les dades pujades s'ha de crear un flux des de la pestanya "Flow". Es crea posant-li un nom i seleccionant els sets de dades que es volen transformar.



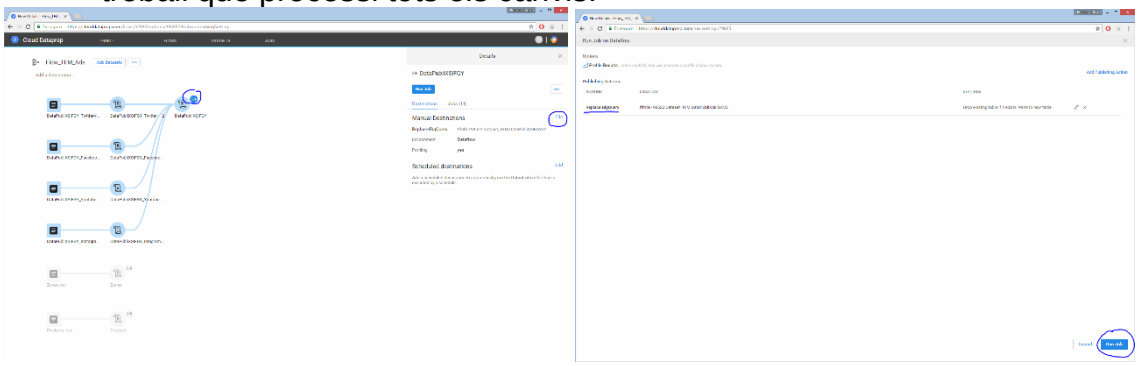
5. Marcant un set de dades afegit al flux poden crear una recepta de canvis.



6. A la recepta de canvis creada s'afegeixen les passes editant-la.



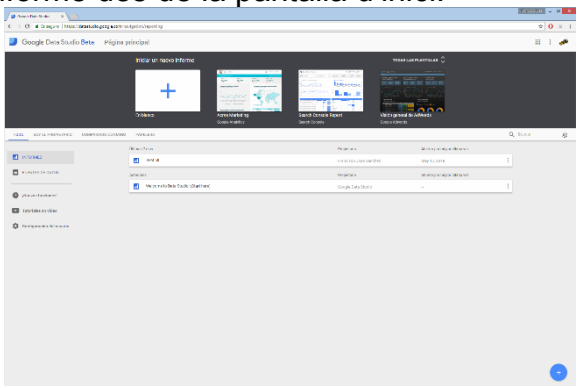
7. Per finalitzar és crea una destinació de les dades transformades i un treball que processi tots els canvis.



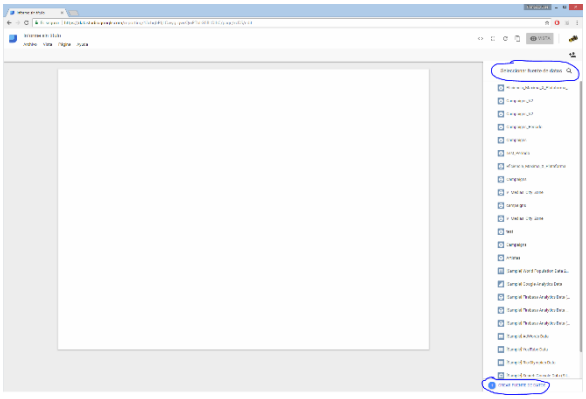
8. La destinació s'indicarà directament Bigquery

C. Creació d'informes amb Data Studio

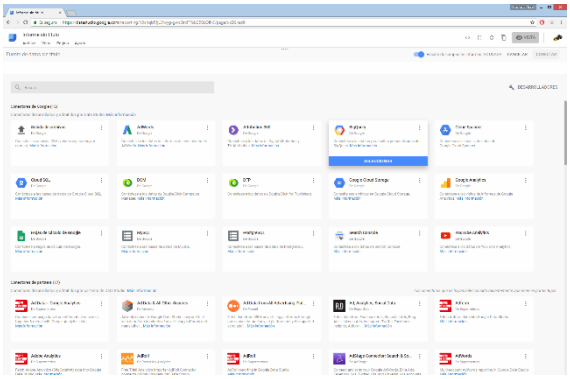
1. Es crea un informe des de la pantalla d'inici.

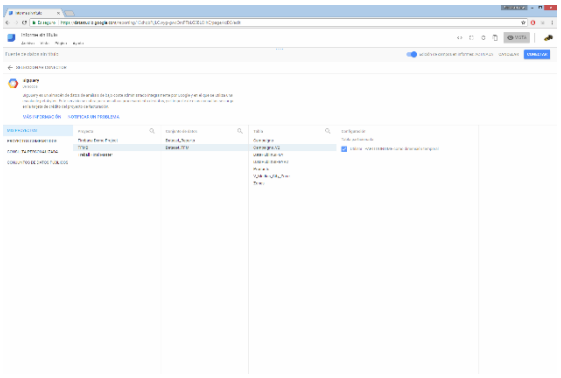


2. A continuació es selecciona les fonts de dades o es creen de noves si no existeixen

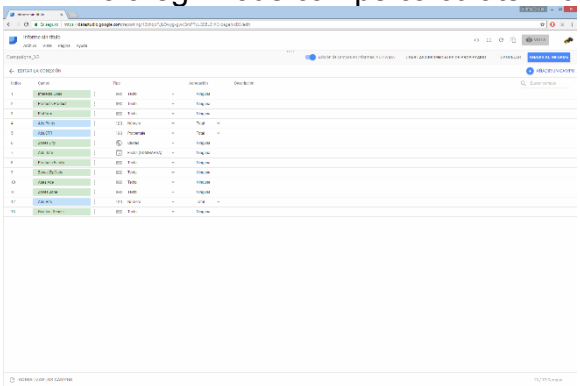


3. Seleccionem de origen "Bigquery" i després la taula o vista d'on volem fer l'informe.





4. Data Studio detecta les mètriques i les dimensions però es pot modificar o afegir nous camps calculats. Pulsem "Añadir Al informre".



5. Finalment es creen els informes, seleccionant els gràfics que es volen afegir, configurant-los amb les seves dimensions i les seves mètriques.