Universitat Oberta de Catalunya

uoc.edu

SISTEMA DE INTELIGENCIA DE NEGOCIOS PARA EL ANALISIS DE PUBLIDAD EN ENTORNOS DIGITALES

Keinth Fernández Pérez
Master en Ingeniería Informática
M1.321 – Trabajo final de máster

David Amorós Alcalá
María Isabel Guitart Hormigo

11/06/2018

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

Copyright © AÑO TU-NOMBRE.

Permission is granted to copy, distribute and/or modify this document under the terms of the GNU Free Documentation License, Version 1.3 or any later version published by the Free

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de Inteligencia de Negocios para el análisis de la publicidad en entornos digitales.*
Nombre del autor:	*Keinth Fernández Pérez*
Nombre del consultor/a:	*David Amorós Alcaraz*
Nombre del PRA:	*María Isabel Guitart Hormigo*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Master Ingeniería informática*
Área del Trabajo Final:	*Trabajo Fin de master*
Idioma del trabajo:	*Español*
Palabras clave	*Inteligencia de negocios, publicidad digital, anuncios.*

Resumen del Trabajo:

El presente trabajo final de máster consiste en el diseño e implementación de un sistema de Business Intelligence que facilita la adquisición, almacenamiento y explotación de datos obtenidos durante la publicación de anuncios en las diferentes plataformas digitales como Instagram, Facebook o Youtube, analizando así la eficiencia en la parametrización de los anuncios digitales.

Para la realización del mismo se llevó a cabo una planificación y definición de tareas en donde se seleccionó como metodología Scrum, también se realizó el análisis y selección de las herramientas open source, posteriormente se diseñó e implementó el Data WareHouse y las ETL respectivas y finalmente se implementó un front-end de usuario donde se muestra mediante un dashboard y de forma gráfica e interactiva los indicadores y la analítica de campañas.

Finalmente podemos concluir que el trabajo fue satisfactorio porque se logró el cumplimiento de los objetivos propuestos, así como las respuestas a las diferentes preguntas analísticas del mismo, además fue de gran enriquecimiento profesional debido a los diferentes conocimientos adquiridos en el área de BI.

Abstract:

The present final master's project consists in the design and implementation of a Business Intelligence system that facilitates the acquisition, storage and exploitation of data obtained during the publication of ads on different social media such as Instagram, Facebook or YouTube, analyzing the efficiency in the parameterization of these digital ads.

For the realization, it was carried out a planning and definition of tasks where it was selected as Scrum methodology, the analysis and selection of open source tools was also carried out, after that, the Data Warehouse and the respective ETLs were designed and implemented, and finally a user front end was implemented where the indicators and the campaign analysis are shown graphically and interactively through a dashboard.

Finally, we can conclude that the work was satisfactory because the proposed objectives were achieved, as well as the answers to the different analytical questions, and it was also a great professional enrichment due to the different knowledge acquired in the BI area.

Índice

Lista de figuras

Lista de Tablas

1. Introducción

1.1 Contexto y justificación del Trabajo

En la actualidad la digitalización impacta todos los aspectos del negocio tanto así que todas las empresas sin importar su tamaño han tenido que buscar presencia en internet a través del marketing digital, mediante el cual han logrado tener muchos beneficios como fidelización de clientes, posicionamiento de la marca, entendimiento de los cambios de comportamiento del consumidor, mejoras en la comercialización de los productos y el posicionamiento global del negocio.

Sin embargo, al implementar una estrategia de marketing, es importante que esta sea efectiva y para esto es necesario llevar la medida del éxito, es decir se requiere el uso de indicadores y estadísticas para medir y evaluar el objetivo trazado, al igual que su progreso y si es necesario realizar los ajustes necesarios durante su desarrollo, todo esto en el menor tiempo posible y con la mayor facilidad para ser compartidas.

Surge entonces la pregunta: cómo lo hacemos? y la respuesta es muy sencilla a través de sistemas BI, los cuales nos permiten transformar los datos en información y a su vez esa información en conocimiento, que finalmente beneficiara la toma de decisiones para el negocio.

“La Inteligencia de Negocio actúa como un factor estratégico para una empresa u organización, generando una potencial ventaja competitiva, que no es otra que proporcionar información privilegiada para responder a los problemas de negocio: entrada a nuevos mercados, promociones u ofertas de productos, eliminación de islas de información, control financiero, optimización de costes, planificación de la producción, análisis de perfiles de clientes, rentabilidad de un producto concreto, entre otros” [1].

El presente trabajo busca diseñar e implementar un entorno Business Intelligence (BI) que permita analizar la eficiencia en la parametrización de anuncios en las principales plataformas digitales, analizando mediante cubos de datos la información y determinando que variaciones sobre la parametrización provocan mejores resultados, además de si existe una relación con la tipología de anuncio.

1.2 Objetivos del Trabajo

Objetivo General: Diseño e implementación de un sistema de Business Intelligence que facilite la adquisición, el almacenamiento y la explotación de datos obtenidos durante la publicación de anuncios en plataforma digitales como Instagram, Facebook o Youtube analizando así la eficiencia en la parametrización de anuncios en las principales plataformas digitales.

Objetivos específicos:

- Diseñar un almacén de datos (Data Warehouse) que permita almacenar la información adquirida en los diferentes orígenes.

- Implementar este almacén de datos y programar los procesos ETL (extracción, transformación y carga) que permitan alimentar el DWH a partir de los ficheros base facilitados.

- Analizar las diferentes plataformas BI OS disponibles en el mercado que nos permitan explorar la información almacenada.

- Elegir una de estas herramientas de tal forma que se disponga de una capa de software para el análisis de la información.

Al final del proyecto se pretende resolver los siguientes interrogantes:

- ¿Qué regiones o ciudades tienen mejores indicadores de efectividad? ¿Hay alguna relación con el producto o familia de productos?
- ¿Existen una relación entre la mejora de los indicadores de efectividad con algún segmento de la población objetivo?
- ¿Hay alguna plataforma donde, bajo las mismas condiciones, se obtengan mejores tasas de visualización?
- ¿Existen relaciones entre plataformas y franjas de edad de usuarios que provoquen mejores tasas de visualización?
- ¿El conocimiento del grupo de interés de los usuarios podría ayudar a mejorar los indicadores para determinados productos?

1.3 Enfoque y método seguido

Las empresas modernas dedicadas al marketing digital, desean tecnología de fácil adopción, que promueva la velocidad y la flexibilidad de los procesos de sus áreas estratégicas, así como de su cadena productiva.

Bajo este contexto y teniendo en cuenta la complejidad natural de un proyecto para la implementación de una solución integral de Business Intelligence, se empleará en la gestión del proyecto la metodología ágil **SCRUM**, que ofrece un marco de trabajo por medio del cual se pueden abordar proyectos complejos y adaptativos, a la vez que se entregan productos de gran valor productivo y creativo, de forma progresiva y ordenada [2].

Teniendo en cuenta este marco de desarrollo incremental e iterativo, el proyecto inicia con una reunión de planificación (***Sprint Planning Meeting***) liderada por el ***Scrum Master*** (para efectos del presente proyecto, yo) y el ***Scrum Team*** (en este caso, complementado por el tutor y el validador del proyecto) donde se estable un plan de trabajo (sección 1.4) a partir de una idea clara de lo que deseamos lograr (***Story Telling***: BI para análisis de publicidad en entornos digitales). Este plan se compone de varios entregables (***Sprints***),

que no son más que los paquetes de trabajo alrededor de un entregable, los cuales se llevaran a cabo en un tiempo corto no superior a 30 días.

Cada entregable (***Sprint***), tendrá una definición de qué se va a construir, un diseño y un plan flexible que guiará su construcción, las actividades a realizar y el producto (entregable) resultante, para efectos del proyecto cada Sprint corresponde a las entregas de las PEC (sección 1.4).

Se hará seguimiento a través de una reunión mensual (***Daily Scrum***) del ***Scrum Master*** con el equipo de desarrollo (para efectos del proyecto, yo misma) y revisiones por demanda (***Sprint Review***) del ***Scrum Master***, el equipo de desarrollo (**Scrum Team**), el dueño del producto (***Product Owner***) y los interesados, donde se validaran el avance en el desarrollo de los elementos que componen el entregable de este Sprint, más el plan para terminarlos, y se registraran en la lista de pendientes del Sprint (***Sprint Backlog***). El Script Backlog dictará el nivel de avance del proyecto y los elementos que tendrán que refinarse, en caso de ser necesario, dentro de las actividades propuestas de los siguientes sprints.

Marco de referencia para el diseño del almacén de datos

Para la concepción y el diseño de los procesos ETL y el Data Warehouse, piezas centrales del Sistema BI que implementaremos, se tomará como marco de referencia teórico las definiciones de Bill Inmon (https://en.wikipedia.org/wiki/Bill_Inmon) y Ralph Kimball (https://en.wikipedia.org/wiki/Ralph_Kimball).

Son teorías distintas que persiguen los mismos objetivos alrededor del diseño de un Data Warehouse. “Kimball representa la relación de los datos con el usuario final, la flexibilidad y la rápida explotación de la información. Por otra parte, Inmon representa la pulcritud en el diseño y el respeto por una serie de normas que garanticen la exactitud de los datos, su integración y su coherencia”[3].

“Con el tiempo, los profesionales del Data Warehousing han aprendido a fusionar ambos marcos en modelos híbridos que recogen lo mejor de ambas teorías, con estructuras que complementan muy bien el ciclo de vida de la información mediante procesos ETL y a su vez modelos dimensionales para representar la información de una forma que facilite su explotación y analítica” [4].

Marco de referencia para el diseño del cuadro de mando

En la actualidad los activos intangibles, cómo los datos, son la fuente más importante de ventaja competitiva de las empresas de cualquier sector, se necesitan herramientas que describan los beneficios y las oportunidades basadas en el conocimiento y las estrategias de creación de valor en función de una mejor información. Sin estas herramientas, las empresas tendrán dificultades para gestionar lo que no puedan describir o medir.

Esto resulta muy familiar en cualquier empresa, pero sobre todo en aquellos sectores con una gran dinámica como el sector de la publicidad, dónde se necesita de un lenguaje común para comunicar la estrategia digital, así como procesos y sistemas que las ayuden a implementarla y obtener información o respuesta sobre ella, de ser posible en tiempo real. Ese lenguaje común es lo que constituye el **cuadro de mando** como herramienta clave dentro de la estrategia de explotación que abordaremos en este proyecto.

Para la concepción y el diseño del cuadro de mando, se tomará como marco de referencia teórico las definiciones que Robert Kaplan y David Norton presentaron en el número de enero/febrero de 1992 de la revista Harvard Business Review, titulado "The Balanced Scorecard: Translating Strategy Into Action".

Sus autores plantean (Kaplan, Robert S. y David P. Norton, The Balanced Scorecard: Translating Strategy Into Action, Boston, MA: Harvard Business School Press, 1996 https://www.leadersummaries.com/ver-resumen/como-utilizar-el-cuadro-de-mando-integral) que gestionar una empresa teniendo en cuenta solamente los indicadores financieros tradicionales (stock, ingresos, gastos, etc.) hace perder de vista la importancia de los activos intangibles de una empresa como fuente principal de ventaja competitiva, así se logra ampliar los objetivos de mejora y desarrollo a aspectos tan relevantes como lo son las relaciones con los clientes, las habilidades y motivaciones de los equipos de trabajo, la cadena de suministro, la innovación, entre otros

De ahí surge la necesidad de crear una nueva metodología para medir las actividades de una compañía en términos de su visión y estrategia, proporcionando a los gerentes una mirada global del desempeño del negocio, conocida como Balanced Score Card o Cuadro de Mando Integral (CMI)[5].

El CMI es una herramienta de administración de empresas, que mucho más allá que del sencillo diseño de una visualización con propósito específico, además dicta un método para medir continuamente cuándo una compañía y sus equipos de trabajo alcanzan los resultados definidos por el plan estratégico. Adicionalmente, un sistema como el CMI permite detectar las desviaciones del plan estratégico y expresar los objetivos e iniciativas necesarios para reconducir la situación.

Algunos de estos conceptos los abordaremos, dentro del presente proyecto, al momento de definir los indicadores de publicidad enfocados en marketing sobre canales digitales y en el diseño e implementación del cuadro de mando interactivo empleado para la visualización de estos indicadores.

1.4 Planificación del Trabajo

A continuación se detallan los diferentes recursos que constituyen este proyecto:

- Humano: Persona que desarrolla el TFM y que aporta su tiempo, esfuerzo y dedicación para la realización del mismo.
- Material y/o Físico: Computadora empleada para el desarrollo del proyecto.
- Tecnológicos: La herramienta BI seleccionada para el desarrollo del proyecto y la base de datos.
- Documentales: Diferentes libros electrónicos, blog, páginas de internet, etc., que servirán de apoyo para la investigación y desarrollo del proyecto.
- Financieros: el desarrollo del proyecto genero gastos en servicios públicos (energía eléctrica).

Se definió además la planificación inicial del trabajo con base en las fases e hitos definidos para el cumplimiento de cada uno de los objetivos y los tiempos de entregas de las PEC, estableciendo finalmente un calendario de lunes a viernes para la realización del proyecto, a continuación el listado de hitos los cuales se resaltan en amarillo:

Nombre de tarea	Duración	Comienzo	Fin
Sistema BI para el análisis de la publicidad en entornos	**76 días**	**lun 26/02/18**	**lun 11/06/18**
Planificación del Proyecto	6 días	lun 26/02/18	lun 05/03/18
Entrega PEC1 - Planificación Proyecto	**0 días**	**lun 05/03/18**	**lun 05/03/18**
Análisis y Selección de las herramientas	11 días	mar 06/03/18	mar 20/03/18
Creación del Entorno de Trabajo	3 días	mié 21/03/18	vie 23/03/18
Diseño del Datawarehouse	11 días	lun 26/03/18	lun 09/04/18
Entregable PEC2 - Selección Herramienta BI y Diseño DWH	**0 días**	**lun 09/04/18**	**lun 09/04/18**
Construcción del Datawarehouse	4 días	mar 10/04/18	vie 13/04/18
Diseño ETL	10 días	lun 16/04/18	vie 27/04/18
Construcción ETL	5 días	lun 30/04/18	vie 04/05/18
Entregable PEC3 - Diseño y Construcción ETL	**0 días**	**lun 07/05/18**	**lun 07/05/18**
Implementación del Front-End del usuario	12 días	mar 08/05/18	mié 23/05/18
Análisis Resultados Obtenidos	8 días	mié 23/05/18	vie 01/06/18
Conclusiones	5 días	lun 04/06/18	vie 08/06/18
Entrega Final del Proyecto	**0 días**	**lun 11/06/18**	**lun 11/06/18**

Tabla 1. Hitos del proyecto

A continuación se muestra a través del diagrama de Gantt todas las fases, tareas e hitos establecidos que se realizaran para el desarrollo del proyecto y el tiempo de cada una de ellas:

Id	Nombre de tarea	Duración	Comienzo	Fin
1	**Sistema BI para el análisis de la publicidad en entornos**	76 días	lun 26/02/18	lun 11/06/18
2	**Planificación del Proyecto**	**6 días**	**lun 26/02/18**	**lun 05/03/18**
3	Establecer Contexto y Justificación	1 día	lun 26/02/18	lun 26/02/18
4	Definición de objetivos	1 día	mar 27/02/18	mar 27/02/18
5	Descomposición del trabajo en tareas	1 día	mié 28/02/18	mié 28/02/18
6	Planificación Tareas en Hitos	3 días	jue 01/03/18	lun 05/03/18
7	Entrega PEC1 - Planificación Proyecto	0 días	lun 05/03/18	lun 05/03/18
8	**Análisis y Selección de las herramientas**	**11 días**	**mar 06/03/18**	**mar 20/03/18**
9	Análisis de las herramientas BI	5 días	mar 06/03/18	lun 12/03/18
10	Documentación Memoria	3 días	mar 13/03/18	jue 15/03/18
11	Selección Herramienta para el proyecto	2 días	vie 16/03/18	lun 19/03/18
12	Propuesta Herramienta seleccionada	1 día	mar 20/03/18	mar 20/03/18
13	**Creación del Entorno de Trabajo**	**3 días**	**mié 21/03/18**	**vie 23/03/18**
14	Verificación requisitos de instalación	1 día	mié 21/03/18	mié 21/03/18
15	Instalación Servidor Base de Datos	1 día	jue 22/03/18	jue 22/03/18
16	Instalación del Software BI	1 día	vie 23/03/18	vie 23/03/18
17	**Diseño del Datawarehouse**	**11 días**	**lun 26/03/18**	**lun 09/04/18**
18	Análisis de requerimientos	2 días	lun 26/03/18	mar 27/03/18
19	Modelo Conceptual de Datos	3 días	mié 28/03/18	vie 30/03/18
20	Modelo Lógico de Datos	3 días	lun 02/04/18	mié 04/04/18
21	Modelo Fisico de Datos	3 días	jue 05/04/18	lun 09/04/18
22	Entregable PEC2 - Selección Herramienta BI y Diseño DWH	0 días	lun 09/04/18	lun 09/04/18
23	**Construcción del Datawarehouse**	**4 días**	mar 10/04/18	vie 13/04/18
24	**Diseño ETL**	**10 días**	lun 16/04/18	vie 27/04/18
25	**Construcción ETL**	**5 días**	lun 30/04/18	vie 04/05/18
26	Entregable PEC3 - Diseño y Construcción ETL	0 días	lun 07/05/18	lun 07/05/18
27	**Implementación del Front-End del usuario**	12 días	mar 08/05/18	mié 23/05/18
28	**Análisis Resultados Obtenidos**	8 días	mié 23/05/18	vie 01/06/18
29	**Conclusiones**	5 días	lun 04/06/18	vie 08/06/18
30	Entrega Final del Proyecto	0 días	lun 11/06/18	lun 11/06/18

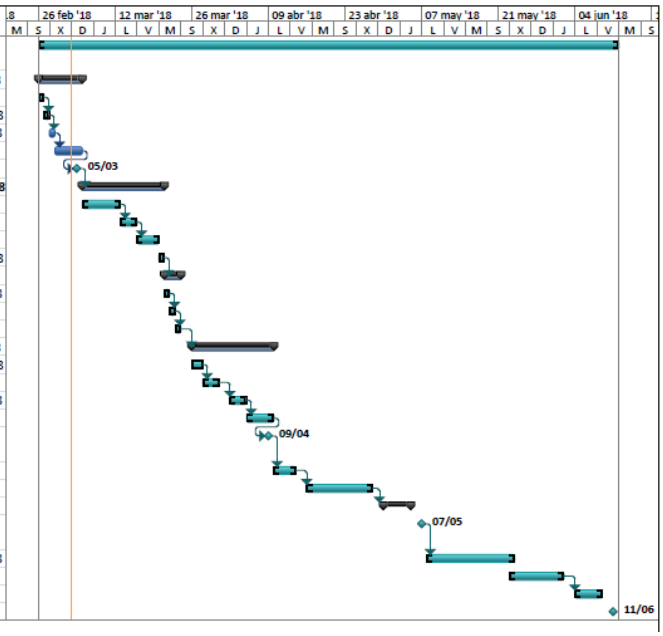


Figura 1. Diagrama de Gantt de la Planificación del Proyecto

1.5 Breve sumario de productos obtenidos

El producto resultante será un sistema de Business Intelligence para el análisis de campañas publicitarias realizadas a través de canales digitales.

A lo largo de la ejecución del proyecto se generarán entregables para los diferentes componentes del sistema:

- Un motor de integración para la extracción, procesamiento y transformación de datos (ETL) a partir de un conjunto de datos (datasets) en Excel, que pueden ser el resultado de una exportación de datos mediante la integración a través de API Rest o mecanismos similares provistos por las plataformas digitales donde se despliegan las campañas.

- Una bodega de datos (Data Warehouse) con una estructura suavizada y ajustada al análisis de campañas de marketing digital.

- Y finalmente, un front-end de usuario (cuadro de mando) que será la capa de visualización donde se mostrará mediante un dashboard y de forma gráfica e interactiva los indicadores y la analítica de campañas.

1.6 Breve descripción de los otros capítulos de la memoria

Teniendo en cuenta la planificación realizada, el proyecto consta de 7 capítulos:

- Capítulo 1: Detalla el problema planteado en el proyecto, los objetivos y la planificación de las tareas definidas en hitos.

- Capítulo 2: Corresponde al marco teórico sobre lo que es el Bussines Intelligence, así como el análisis y selección de las herramientas BI a implementar.

- Capítulo 3: Define el diseño e implementación del DWH, se realiza una análisis y comprensión de los datos y el diseño del modelo conceptual, lógico y físico de los datos.

- Capítulo 4: Corresponde al diseño e implementación de la ETL, se explica el diseño e implementación de los distintos procesos ETL generados para la carga de datos en el DWH.

- Capítulo 5: Detalla la creación del Entorno BI así como la implementación del mismo.

- Capítulo 6: Corresponde al análisis de los datos y conclusiones, se muestran las visualizaciones, las analíticas de datos y la unión de todos estos elementos en un cuadro de mando.

- Capítulo 7: Contiene las conclusiones generales del trabajo de grado.

- Capítulo 8,9 y 10: Corresponden al Glosario, bibliografía y anexo respectivamente.

2. Análisis y selección de la Herramienta BI

2.1 Arquitectura de un sistema de Business Intelligence

Cuando hablemos de Business Intelligence, hay que pensar en una fábrica de información donde se integran un conjunto de conceptos, de procesos y de datos por medio de herramientas tecnológicas, cuyo producto resultante es nuevo conocimiento y mejor información, para la toma de decisiones. Y aunque Business Intelligence, no implica tecnología en sí mismo, si está soportado por un ecosistema de herramientas tecnológicas y mecanismos de automatización que facilitan la explotación de la información por parte de los interesados.

La existencia de este ecosistema de herramientas tecnológicas especializadas en BI supone, entre otros, los siguientes beneficios:

- Disponer un entorno para la minería, el reporting y la analítica de datos
- Estructurar un gobierno de datos
- Propender por una calidad de datos
- Contar con un medio propicio para la democratización de los datos

Desglosando un ambiente de BI en sus partes más esenciales, nos encontramos con una conformación mediante capas, siendo la primera de ellas la capa de orígenes de datos constituida por los datos almacenados en los diferentes sistemas de información transaccionales/operacionales internos y externos a la organización. Seguida por la capa de integración donde estos datos son extraídos, transformados en otro tipo de datos más acordes para la analítica o el diseño de indicadores, y finalmente cargados a una bodega de datos, todo esto mediante motores especializados de ETL (extract, transform and load). El resultado de estos procesos de integración, serán repositorios de datos más pequeños, configurados para la consulta y estructurados mediante etiquetas (metadatos), lo que conoceremos como la capa de la bodega de datos (Data Warehouse). Por último, teniendo los datos estructurados se dispondrá la explotación de estos a través de la capa de aplicación o presentación, donde se emplearán herramientas especializadas para el reporting, la analítica y visualización de datos.

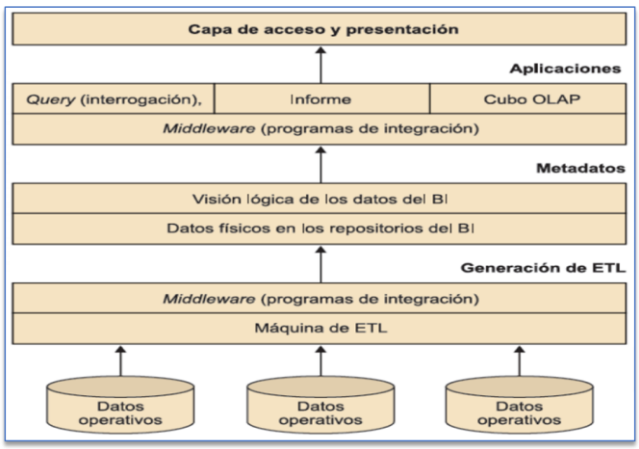


Figura 2. Business Intelligence Roadmap: Moss y Atre (2003).

Estas tres capas serán incluidas en el análisis de herramientas (secciones 2.3, 2.4 y 2.5) para determinar la escogencia (sección 2.6), bajo criterios técnicos y funcionales, de la arquitectura definitiva con la que se realizará la implementación del sistema BI.

2.2 El Business Intelligence Open Source (BI OS)

En los últimos 10 años estamos observando cómo los grandes fabricantes están cambiando sus estrategias para crear productos, servicios y los métodos de innovación. Algunos adquieren proyectos de software libre altamente posicionados («Top 50 Open Source Companies: Where Are They Now?». http://www.channelfutures.com/open-source/top-50-open-source-companies-where-are-they-now [en inglés]. Fechado: 29/04/2017.) y otros optan por abrir bajo la misma figura el código fuente de sus productos, por ejemplo, a través de API's.

Aquellos fabricantes que han liberado su código fuente han evidenciado como su software ha sido ampliamente aceptado por una comunidad activa y dinámica de desarrolladores, que han generado una documentación densa a su alrededor y han indicado un panorama claro de sostenimiento al producto en el tiempo. A esto se le suma, el hecho de ser tecnologías adoptadas por algunos de los principales fabricantes de productos y servicios de software, así como, sus principales canales afiliados de distribución y soporte, lo que supone un atractivo para las empresas que buscan adoptar este tipo de tecnologías.
Sin hablar específicamente del nicho de inteligencia de negocios, estudios recientes realizados a una base de cerca de 250 empresas alrededor de Latinoamérica hablan de que el 70% de las compañías ya usan herramientas open source para apoyar sus procesos corporativos, y mejor aún, otro 8% piensan adoptarlas en los siguientes 12 meses, lo que supone una madurez confirmada en el ciclo de adopción de este tipo de soluciones («Estatus de adopción de cloud y open source en Latinoamérica». https://sg.com.mx/buzz/estatus-adopci-n-cloud-y-open-source-latinoam-rica [en español]. Fechado: 25/04/2017).

Ahora, teniendo en cuenta que la economía mundial parece aún no salir de una crisis que afecta a muchos sectores, los presupuestos limitados de inversión y gasto, es un factor relevante hoy al interior de las empresas, por lo cual acceder a tecnología especializada bajo un esquema de reducción del costo de adquisición (TCO) es una realidad y hace parte de la estrategia financiera adoptada por las diferentes compañías.

Adicional, y pensando en el sector de servicios de publicidad y mercadeo, la alta competencia exige una dinámica mayor dentro de los procesos de innovación al interior de las empresas que ofrecen servicios de marketing digital, lo que supone probar nuevas estrategias en un menor tiempo y a un menor costo.

Por último, los principales proyectos alrededor de los 3 mercados que componen la inteligencia de negocios (BI, Business Analytic y Big Data) tienen su centro o su punta de lanza, en la comunidad open source.

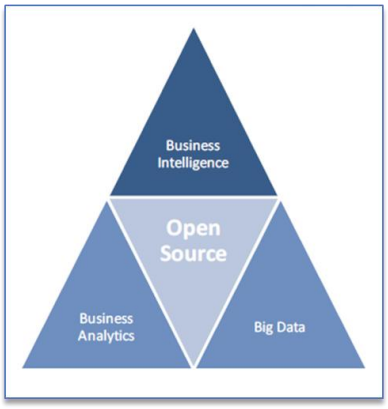


Figura 3. Impacto del Open Source en el ecosistema BI.

Todo lo anteriormente argumentado se convierte en suficiente sustento para considerar dentro del proyecto la valoración y escogencia de herramientas BI Open Source. A continuación, analizaremos las principales opciones open source dentro de cada uno de los componentes de un BI: integración, bodegas y visualización de datos.

2.3. Metodología para el análisis y evaluación de las herramientas

Para la realización de la evaluación de la herramienta, se han definido unas plantillas que facilitaran la comparación de características, el análisis de pros/contras y la calificación de los aspectos relevantes dentro de un ecosistema de inteligencia de negocios entre herramientas, entre las diferentes herramientas.

Para la comparación de características, se manejarán un listado de categorías, que deberán ser detalladas brevemente para cada herramienta en cada una de las tres capas: integración, bodega de datos y visualización.

Capa de integración de datos	
CATEGORÍA	ASPECTOS RELEVANTES
Modelamiento	
Extracción y carga	
Transformación	
Trazabilidad	
Capa bodega de datos	
CATEGORÍA	ASPECTOS RELEVANTES
Escalabilidad	
Rendimiento	
Seguridad	
Administración	
Capa de visualización de datos	
CATEGORÍA	ASPECTOS RELEVANTES
Conexión a datos	
Descubrimiento de datos	
Colaboración	
Administración de datos	

Tabla 2. Categorías a evaluar para la selección de la herramienta BI.

Para el análisis de puntos a favor y en contra, se desglosarán las categorías en un listado de características técnicas, las cuales se les asignará un peso que determinará su importancia y las cuales deberán calificarse según su nivel de madurez empleando la escala definida en la matriz de cumplimiento, para cada herramienta en cada una de las tres capas: integración, bodega de datos y visualización.

Matriz de cumplimiento

OPCIÓN	CRITERIO DE MADUREZ	VALOR
No cumple	Madurez nula, no cumple la característica técnica	0
Bajo	Nivel bajo de madurez, característica en desarrollo	1
Medio	Nivel medio de madurez, característica incorporada parcialmente	2
Alto	Nivel alto de madurez, característica incorporada y validada	3

Tabla 3. Matriz de cumplimiento para las herramientas BI.

Capa de integración de datos

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA
Modelamiento	Diseño visual de procesos ETL	7
	Interfaz acoplada e intuitiva	1
Extracción y carga	Extracción y carga de datos completa o incremental	3
	Multiples conectores a fuentes remotas y archivos	8
Transformación	Validación de consistencia e integridad de datos	6
	Procesamiento paralelo de grandes volumenes de datos	5
Trazabilidad	Alertas cuando se detectan errores en los flujos ETL	2
	Identificación de cuellos de botella en los flujos ETL	4

Capa bodega de datos

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA
Escalabilidad	Soporte a grandes volumenes de datos	8
	Adopción por otros fabricantes para adaptaciones dirigidas a DWH	1
Rendimiento	Manejo de concurrencia en lectura y escritura	7
	Soporta particionamiento de tablas	6
Seguridad	Permite operaciones de lectura mientras se realizan tareas de mantenimiento	4
	Altamente soportada por la comunidad de desarrolladores	5
Administración	Administración intuitiva de las bases de datos multidimensionales	2
	Flexibilidad para crear y dar soporte a nuevos tipos de datos	3

Capa de visualización de datos

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA
Conexión a datos	Conexión a fuentes de datos del tipo DWH	8
	Conexión simultanea a varias fuentes de datos	7
Descubrimiento de datos	Metodos de exploración en profundidad (drill-down)	6
	Cuadros de mando avanzados con riqueza visual e interactivos	5
Colaboración	Publicación de dashboards en un Cloud BI	3
	Herramientas para compartir y colaborar	2
Administración	Auto-servicio y preparación de datos	4
	Facilidad para escalar a un versión premium	1

Tabla 4. Características técnicas a evaluar para la selección de las herramientas BI.

Finalmente, la calificación de cada herramienta estará determinada por el promedio de las calificaciones de cada característica, que se calcularan multiplicando el grado de madurez de la característica por su escala de importancia dentro de un ecosistema BI. Las herramientas con el mejor score de calificación total serán elegidas para componer la arquitectura BI que se implementará.

Capa de integración de datos				
CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Modelamiento	Diseño visual de procesos ETL	7		
	Interfaz acoplada e intuitiva	1		
Extracción y carga	Extracción y carga de datos completa o incremental	3		
	Multiples conectores a fuentes remotas y archivos	8		
Transformación	Validación de consistencia e integridad de datos	6		
	Procesamiento paralelo de grandes volumenes de datos	5		
Trazabilidad	Alertas cuando se detectan errores en los flujos ETL	2		
	Identificación de cuellos de botella en los flujos ETL	4		
			TOTAL	

Tabla 5. Calificación final para la selección de las herramientas BI.

2.4. Análisis y evaluación de herramientas para la capa de integración de datos

Como parte de la capa para la integración de datos, están los procesos ETL, gracias a los cuales es posible:

- Mover los datos desde una o múltiples fuentes.
- Reformatear esos datos y depurarlos, si es necesario.
- Cargarlos en un repositorio estructurado sea este una base de datos, un Data Mart o un Data Warehouse.
- Una vez alojados en el repositorio destino, analizarlos.

A continuación, analizaremos tres alternativas de herramientas open source especializadas en integración de datos (ETL):

Pentaho Data Integration

Solución tecnológica conocida como Kettel, es una de las herramientas ETL de código abierto más potentes y versátiles a la hora de diseñar los procesos de integración enfocados en el poblamiento de datos del Data Warehouse.

Como herramienta ETL, es visual y además permite la conexión a diversas fuentes y arquitecturas para finalmente cargar los datos en un repositorio, entre otras funciones, como: sincronizar, enmascarar y migrar datos entre diferentes aplicaciones.

Según Gartner, Pentaho es una herramienta que destaca dentro del grupo de líderes de ETL, incorporando características de conectividad, capacidad de entrega de datos, de metadatos y modelados de datos, de diseño y entorno de desarrollo, de gestión de datos de administración, además de capacidades SOA y un cierto grado de compactación, consistencia e interoperabilidad. Se incluyó por primera vez en el informe de Gartner del año 2009 como software open source de probada eficacia, manteniendo su presencia dentro de este estudio en la actualidad [6].

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Modelamiento	• Aunque cuenta con herramienta para diseño visual de ETL, algunos aspectos relacionados con la transformación de los datos requiere un conocimiento de SQL y programación. • Es modular y la Suite es muy completa, pero los módulos vienen desacoplados de forma nativa y hay que invertir un esfuerzo importante por integrarlos, lo que condiciona el uso y explotación de las funcionalidades para el modelamiento de flujos.
Extracción y carga	• La carga incremental no viene incorporada de forma nativa, hay opciones de programarla usando diferentes técnicas. • Ofrece una amplia variedad de conectores a fuentes de datos de ficheros, bases de datos y hasta entornos de big data.
Transformación	• Incorpora una funcionalidad llamada Data Validator que ayuda en la validación de tipologías e integridad entre fuentes, transformación y destino. • El paralelismo es muy sencillo de realizar en procesos con grandes volumenes de información, empleando la opción Distribute Data.
Trazabilidad	• El manejo de errores se hace a nivel de proceso y el control de flujos es muy limitado. Aspecto mejorable. • Incorpora diferentes niveles de logs, suficiente para analizar las ejecuciones de los procesos ETL.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Modelamiento	Diseño visual de procesos ETL	7	2	14
	Interfaz acoplada e intuitiva	1	1	1
Extracción y carga	Extracción y carga de datos completa o incremental	3	1	3
	Multiples conectores a fuentes remotas y archivos	8	3	24
Transformación	Validación de consistencia e integridad de datos	6	3	18
	Procesamiento paralelo de grandes volumenes de datos	5	3	15
Trazabilidad	Alertas cuando se detectan errores en los flujos ETL	2	1	2
	Identificación de cuellos de botella en los flujos ETL	4	3	12
			TOTAL	**11,1**

Tabla 6. Análisis y evaluación de la herramienta Pentaho Data Integration

Talend Open Studio

Desarrollada en Java, Talend tiene una versión de pago y otra comunitaria distribuida como software libre a la que llama Talend Open Studio. Es precisamente su versión comunitaria la que ha propagado su adopción permitiendo aprender y probar la herramienta con buena parte de sus funcionalidades incorporadas.

Talend permite realizar modelamiento de procesos ETL de forma visual y sencilla. De todas las soluciones open source para la integración de datos, Talend es la más potente en procesos de extracción, transformación y carga.Talend básicamente extrae los datos desde diferentes fuentes que pueden ser base de datos, archivos, aplicaciones, web services, correo electrónico, etc. Posteriormente se pueden aplicar diferentes tipos de transformaciones a los datos mediante mecanismos de join, lookup, duplicación, hacer cálculos, y todo previo al paso final hacia el Data Warehouse.

Evaluación de la herramienta

CATEGORÍA	ASPECTOS RELEVANTES
Modelamiento	• Interfaz gráfica con funcionalidad de arrastrar y soltar para la diagramación de procesos ETL. • Interfaz de usuario unificada en todos los componentes, que requiere un curva muy corta de aprendiaje.
Extracción y carga	• Puede realizar carga incremental de datos teniendo como apoyo una campo de la fuente de datos. • Talend cuenta con una gran cantidad de conectores para integrarse con bases de datos, con ficheros de diferentes tipos (XML, Excel, CSV, TSV, JSON, etc), y a servicios/aplicaciones concretas como: SAP, SugarCRM, OpenBravo, SalesForce, Alfresco, entre otras relacionadas con el ecosistema Big Data.
Transformación	• Gran performance en calculo de agregaciones y lookups. • Paralelismo muy reducido en las versión gratuita. Funcionalidad avanzada en las versión premium (Integration Suite).
Trazabilidad	• El manejo de errores se hace a nivel de logs. Adicional, el control del flujo se hace a nivel de row, iterate o row lookup y además se usan disparadores para control de ejecución y orquestación de procesos. • Talend incluye herramientas del tipo Data Profiling para identificar los cuellos de botella. Adicional, cuenta con funcionalidad de log que se pueden registrar en base de datos, consola y archivos, distinguiendo la información de estadísticas, de métricas de procesos y trazas de error.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Modelamiento	Diseño visual de procesos ETL	7	3	21
	Interfaz acoplada e intuitiva	1	3	3
Extracción y carga	Extracción y carga de datos completa o incremental	3	3	9
	Multiples conectores a fuentes remotas y archivos	8	3	24
Transformación	Validación de consistencia e integridad de datos	6	3	18
	Procesamiento paralelo de grandes volumenes de datos	5	2	10
Trazabilidad	Alertas cuando se detectan errores en los flujos ETL	2	3	6
	Identificación de cuellos de botella en los flujos ETL	4	3	12
			TOTAL	**12,9**

Tabla 7. Análisis y evaluación de la herramienta Talend Open Studio

CloverETL Community Edition

Es un software con una suite muy completa que incluye funcionalidades migración, almacenamiento y alimentar otras bases de datos y sistemas especializados del tipo Data Warehouse.

Desde su creación 2002, la plataforma CloverETL se ha basado su desarrollo en tres principios estables: una arquitectura robusta, la creencia que menos es mejor, y la importancia de la planificación a largo plazo [7].

CloverETL está desarrollado en JAVA y de código abierto. Tiene tres presentaciones: mediante línea de comando, aplicación de servidor o incrustado en otras aplicaciones como una librería de JAVA. Además cuenta con un plug-in que agrega una interfaz gráfica para diseñar procesos ETL.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Modelamiento	• La versión comunitaria cuenta con un Designer que facilita la diagramación mediante arrastar y soltar, sólo que está limitada a flujos que no tengan más de 20 bloques de transformación. • La interfaz es intuitiva y acoplada para la fácil apropiación.
Extracción y carga	• La extracción y carga incremental es un recurso incorporado en la ultilidad de Designer habilitada para la versión comunitaria. • La versión comunitaria viene habilitad sólo a un conjunto limitado de bases de datos: MySQL, MS SQL y Oracle. No opera con extracción desde archivos.
Transformación	• La versión comunitaria es muy limitada, con la versión premium se habilitan muchas más opciones, como: la clasificación, los clusters, entre otras. • El paralelismo de datos no es posible en la versión comunitaria, a nivel de segmentación está limitada a la composición del proceso ETL sólo apoyado en máximo 20 bloques y el paralelismo de componente sólo posible en la versión premium.
Trazabilidad	• La orquestación de workflows/jobs con notificaciones de fallos vía email, sólo es posible en la versión premium. • La versión premium, no la comunitaria, cuenta con la utilidad Data Profiler para ver rápidamente conjuntos de datos desconocidos e identificar delays dentro del flujo.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Modelamiento	Diseño visual de procesos ETL	7	2	14
	Interfaz acoplada e intuitiva	1	3	3
Extracción y carga	Extracción y carga de datos completa o incremental	3	3	9
	Multiples conectores a fuentes remotas y archivos	8	1	8
Transformación	Validación de consistencia e integridad de datos	6	1	6
	Procesamiento paralelo de grandes volumenes de datos	5	1	5
Trazabilidad	Alertas cuando se detectan errores en los flujos ETL	2	0	0
	Identificación de cuellos de botella en los flujos ETL	4	0	0
			TOTAL	**5,6**

Tabla 8. Análisis y evaluación de la herramienta CloverETL

2.5. Análisis y evaluación de herramientas para la bodega de datos

La bodega de datos, también conocida en inglés como Data Warehouse, será el repositorio estructurado de datos útiles y orientados a un ámbito específico del negocio, integrado a los diferentes sistemas internos/externos a través de procesos ETL, no volátil, sí variable en el tiempo, y dimensionado para la consulta frecuente, la analítica y la toma de decisiones. Va más allá de la información registrada en los sistemas transaccionales y operacionales, y aunque los datos también residen en una base de datos, está es diseñada para favorecer el análisis y la divulgación eficiente de datos a través de arquitecturas OLAP y procesamiento analítico en línea. Un Data Warehouse contiene comúnmente grandes cantidades de información que se subdividen, en algunos casos, en unidades lógicas más pequeñas conocidas como Data Marts.

A continuación, analizaremos tres alternativas de herramientas open source especializadas en implementación de bodegas de datos (Data Warehouse):

PostgreSQL

Es muy popular, gratuita y de código libre. Ha sido altamente adoptada por grupos empresariales que han desarrollado servicios de almacenamiento en la nube donde PostgreSQL es el core de su arquitectura. Potente y cuenta con una comunidad decente que la respalda, llamada PostgreSQL Global Development Group. También dispone de un gestor propio conocido como PgAdmin y su variante web phpPgAdmin, que reduce la curva de aprendizaje para su administración. Algunos consideran que hoy en día es el único sistema de base de datos realmente libre.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Escalabilidad	• Pensada para soportar grandes volumenes de datos bajo un contexto SQL o noSQL. • Las infraestructuras DBMS ofrecidas por Pivotal y EnterpriseDB, proveedores de soluciones robustas de DWH ubicados dentro del cuadrante mágico de Gartner 2017, están basadas en el código abierto de PostgreSQL.
Rendimiento	• PostgreSQL también proporciona un entorno óptimo en performance, gracias a su método de control de concurrencias multi-versión, que permite leer y escribir de forma simultánea. • Soporta particionamiento de tablas.
Seguridad	• Mediante su tecnología hot-standby permite que los usuarios puedan acceder a las tablas en modo lectura mientras que se realizan los procesos de backup o mantenimiento. • Cuenta con una comunidad decente y muy dinámica que la respalda, llamada PostgreSQL Global Development Group.
Administración	• PostgreSQL cuenta con una interfaz de administración cliente (PgAdmin) y web (phpPgAdmin) muy completa soportada para tratar con volumenes más grandes, consultas más largas y todo lo necesario para administrar un DWH. • Soporta creación de nuevos tipos de datos.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Escalabilidad	Soporte a grandes volumenes de datos	8	3	24
	Adopción por otros fabricantes para adaptaciones dirigidas a DWH	1	3	3
Rendimiento	Manejo de concurrencia en lectura y escritura	7	3	21
	Soporta particionamiento de tablas	6	3	18
Seguridad	Permite operaciones de lectura mientras se realizan tareas de mantenimiento	4	3	12
	Altamente soportada por la comunidad de desarrolladores	5	3	15
Administración	Administración intuitiva de las bases de datos multidimensionales	2	3	6
	Flexibilidad para crear y dar soporte a nuevos tipos de datos	3	3	9
			TOTAL	**13,5**

Tabla 9. Análisis y evaluación de la herramienta PostgreSQL

Oracle MySQL

MySQL es el motor relacional de base de datos más popular entre los desarrolladores, de buen rendimiento y a costo cero, esto último a pesar de ser adquirido por ORACLE.El uso de MySQL en un entorno de Data Warehouse empleando bases de datos multidimensionales puede llevarse a cabo, pero debe contemplar un volumen no elevado de datos, ajustado a las mismas limitaciones técnicas que impone el motor a nivel del tamaño de sus tablas.

Con una correcta implementación que considere el uso de ingenierías de almacenamiento, las caches y las indexaciones disponibles para MySQL, puede ser usada para impulsar una solución OLAP a una fracción del costo de otras alternativas comerciales.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Escalabilidad	• Con grandes volumenes de datos pierde rendimiento. • No tienen referencias de adopción por parte de grandes proveedores de servicios DWH.
Rendimiento	• Le da control a la concurrencia mediante bloqueos pero no la permite de forma natural. • Soporta particionamiento de tablas.
Seguridad	• Para backups en caliente se usa la utilidad mysqldump, que dependiendo del volumen del respaldo puede degradar el rendimiento del motor. • Su soporte está limitado a Oracle, ya no recibe el mismo apoyo de antes, sus desarrolladores más entusiastas han migrado al proyecto MariaDB.
Administración	• Aunque cuenta con interfaz de administración cliente y web, cuentan con funcionalidades muy limitadas para manejar un DWH. • No soporta la creación de nuevos tipos de datos, y su rendimiento es deficiente con consultas recursivas sobre datos jerarquicos.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Escalabilidad	Soporte a grandes volumenes de datos	8	2	16
	Adopción por otros fabricantes para adaptaciones dirigidas a DWH	1	0	0
Rendimiento	Manejo de concurrencia en lectura y escritura	7	1	7
	Soporta particionamiento de tablas	6	3	18
Seguridad	Permite operaciones de lectura mientras se realizan tareas de mantenimiento	4	2	8
	Altamente soportada por la comunidad de desarrolladores	5	2	10
Administración	Administración intuitiva de las bases de datos multidimensionales	2	1	2
	Flexibilidad para crear y dar soporte a nuevos tipos de datos	3	0	0
			TOTAL	**7,6**

Tabla 10. Análisis y evaluación de la herramienta MySQL

MariaDB

Se trata de un fork mejorado de MySQL. El desarrollo de MariaDB surgió como una medida de contingencia de la comunidad desarrolladora de MySQL a la posibilidad de que el producto fuera descontinuada por Oracle, fabricante que se hizo a su propiedad. Los comandos de MySQL, sus interfaces, API e incluso sus librerías se usan también en MariaDB [8], sin embargo, su motor relacional fue completamente modificado para proporcionarle un mayor rendimiento.

“El avance y compromiso de MariaDB es tal que ya ha comenzado a ser adoptado por defecto en algunas distribuciones de Linux muy populares, como por ejemplo CentOS y Fedora, que ya lo trae preinstalado. Como si eso fuera poco, ha sido adoptado dentro de las infraestructuras de gigantes de la tecnología como Google, Wikipedia o Mozilla” [9].

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Escalabilidad	• Con grandes volumenes de datos pierde rendimiento. • No tienen referencias de adopción por parte de grandes proveedores de servicios DWH.
Rendimiento	• Le da manejo a la concurrencia asegurando la Atomicidad, Consistencia, Aislamiento y Durabilidad, en entornos altamente transaccionales, solo que lo hace mejor que MySQL en escenarios altamente transaccionales. • Soporta particionamiento de tablas.
Seguridad	• No se pueden realizar backups o tareas de mantenimiento en caliente. • Recibe todo el soporte de Michael Widenius, su creador y el respaldo de los mejores desarrolladores de la otrora comunidad vinculada al declinado proyecto MySQL AB.
Administración	• Existen muchas interfaces de terceros que se integran a MariaDB aportando el manejo necesario para un DWH. • No soporta la creación de nuevos tipos de datos.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Escalabilidad	Soporte a grandes volumenes de datos	8	2	16
	Adopción por otros fabricantes para adaptaciones dirigidas a DWH	1	0	0
Rendimiento	Manejo de concurrencia en lectura y escritura	7	2	14
	Soporta particionamiento de tablas	6	3	18
Seguridad	Permite operaciones de lectura mientras se realizan tareas de mantenimiento	4	0	0
	Altamente soportada por la comunidad de desarrolladores	5	3	15
Administración	Administración intuitiva de las bases de datos multidimensionales	2	2	4
	Flexibilidad para crear y dar soporte a nuevos tipos de datos	3	0	0
			TOTAL	**8,4**

Tabla 11. Análisis y evaluación de la herramienta MariaBD

2.6. Análisis y evaluación de herramientas para la capa de visualización de datos

Sobre la capa para la visualización de datos, normalmente se disponen de aplicaciones especializadas en la visualización de datos mediante reportes o dashboards. Los dashboards, también conocidos como cuadros de mando, son los preferidos y en esencia permiten la exploración y explotación de los datos de manera intuitiva y visual, a través de una abstracción de los datos mediante representaciones gráficas e indicadores para efectos de analíticas y procesos de toma de decisiones estratégicas en cualquier entorno empresarial.

A continuación, analizaremos tres alternativas de herramientas gratuitas, no necesariamente open source, especializadas en implementación de visualización de datos (dashboard):

Microsoft Power BI

Esta herramienta permite llevar las grillas de Excel y las tablas dinámicas a un nuevo nivel, al de la visualización gráfica e interactiva de datos.

No es en sí misma una herramienta gratuita, pero cuenta una versión desktop que permite realizar todo el trabajo, aunque para compartirlo vía web o mobile se requerirá adquirir la versión paga que adiciona un componente online. La versión gratuita tiene muchas funcionalidades, es muy intuitiva de utilizar y permite limpiar los datos de una manera sencilla. Además, las opciones de conexión a fuentes de datos son muy variadas con mecanismos de refresco en caso de que los datos de origen se modifiquen.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Conexión a datos	• Permite la obtención de datos a partir de ficheros Excel, CSV, XML, JSON, etc, así como, multiples conectores a bases de datos, servicios cloud y plataformas de big data como Hadoop. • Puede trabajar de forma simultanea con varias fuentes.
Descubrimiento de datos	• El mayor atractivo de Power BI es que permite realizar presentaciones interactivas y ajustar las visualizaciones para obtener un mayor detalle de los datos a través de una funcionalidad "drill in". • Facilidad de creación de visualizaciones, basta con arrastrar y soltar los elementos gráficos donde se desee. Incluye paneles, informes y conjuntos de datos que contienen visualizaciones para enriquecer los dashboards.
Colaboración	• Las visualizaciones se pueden publicar online pero se requiere de la versión premium. • La colaboración es amplia en su servicio premium. Se tiene una opción de compartir la visualización entre diferentes usuarios con Power BI que es gratuito.
Administración	• Mantiene el principio de actuación con la que construyo su base de clientes para su otro producto MS-Excel: cinco segundos para descargarla y cinco minutos para sorprender al cliente. Permite transformar y depurar datos provenientes de sus fuentes. • Es muy fácil escalar a la versión Microsoft Power BI Pro pues mantiene un precio muy competitivo y es uno de los más bajos del mercado ($9.99 por usuario, por mes por hasta 10GB/usuario).

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Conexión a datos	Conexión a fuentes de datos del tipo DWH	8	3	24
	Conexión simultanea a varias fuentes de datos	7	3	21
Descubrimiento de datos	Metodos de exploración en profundidad (drill-down)	6	3	18
	Cuadros de mando avanzados con riqueza visual e interactivos	5	3	15
Colaboración	Publicación de dashboards en un Cloud BI	3	1	3
	Capacidadades sociales y de colaboración	2	2	4
Administración	Auto-servicio y preparación de datos	4	3	12
	Facilidad para escalar a un versión premium	1	3	3
			TOTAL	**12,5**

Tabla 12. Análisis y evaluación de la herramienta PowerBI

Tableau Public

En términos de impacto y atractivo en las visualizaciones, Tableu es el gran líder del mercado, muy reconocido por su versión paga, también cuenta con un servicio gratuito denominado Tableau Public. Mientras que la versión paga de Tableau puede conectarse a casí cualquier fuente de datos local/remota, la versión pública se limita a fuentes de datos a través de archivos (p.e. Excel, CSV, JSON, etc), y servicios online tipo OData o Windows Azure Marketplaces y funciona con hasta 1 millón de filas de datos.

Cuenta con un excelente motor de datos in-memory llamada Hyper, enfocada en superar todos los inconvenientes de consulta que pueden presentarse cuando se realizan conexiones en vivo a las fuentes de datos o Data Warehouse. Como su nombre lo indica, para compartir la información se deben publicar los datos en una web de Tableau bajo un perfil público por lo tanto no hay confidencialidad de la información en la versión gratuita.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Conexión a datos	• Conexión a origenes de datos diversos basados en archivos (Excel, CSV, JSON, etc) y a fuentes de datos remotas sólo limitado a ODATA, Google Drive y Tableau Server. • Puede trabajar una visualización conectado a fuentes de datos diversas de forma simultanea.
Descubrimiento de datos	• Opciones avanzadas de análisis, incluyendo líneas de referencia, funciones de predicción y cálculos. • Posibilidad de crear presentaciones interactivas mediante Dashboards e Historias.
Colaboración	• Las visualizaciones se pueden publicar en Tableau Public (cloud BI), aunque no de manera privada. Se tiene una opción de visualización en local mediante Tableau Reader. • La colaboración está limitada a su servicio premium (Tableau Server)
Administración	• Ofrece una plataforma intuitiva, muy visual e interactiva, enfocada en que el flujo de trabajo sea sencillo y sus clientes puedan encontrar fácilmente los insights en sus datos. • El esquema de costos de Tableau Desktop es bastante elevado y fuera del alcance del mercado de PYMES.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Conexión a datos	Conexión a fuentes de datos del tipo DWH	8	1	8
	Conexión simultanea a varias fuentes de datos	7	2	14
Descubrimiento de datos	Metodos de exploración en profundidad (drill-down)	6	3	18
	Cuadros de mando avanzados con riqueza visual e interactivos	5	3	15
Colaboración	Publicación de dashboards en un Cloud BI	3	2	6
	Capacidadades sociales y de colaboración	2	1	2
Administración	Auto-servicio y preparación de datos	4	3	12
	Facilidad para escalar a un versión premium	1	1	1
			TOTAL	**9,5**

Tabla 13. Análisis y evaluación de la herramienta Tableau

Pentaho Community Dashboard

Pentaho incluye dentro de su suite a Pentaho Community Dashboard, para el diseño, desarrollo, edición y visualización de cuadros de mando avanzados.

Para sacar su máximo provecho es necesario instalar todos los componentes que conforman el CTools: CDA (Data Access), CDE (Dashboard Editor), CDF (Dashboard Framework), CCC (Chart Components) y CGG (Graphics Generator). Y aunque es bastante modular, para los efectos de explotación de datos y el diseño de visualizaciones interactivas, hay que invertir un esfuerzo considerable en el alistamiento de la herramienta a nivel de módulos y plug-ins.

Evaluación de la herramienta:

CATEGORÍA	ASPECTOS RELEVANTES
Conexión a datos	• Conexión a origenes de datos diversos, incluyendo los principales entornos DWH y OLAP. • Para lograr la conexión simultanea a varias fuentes de datos de diferente tipo, hay que someter las fuentes a un proceso de union y normalización de datos a través de su módulo Data Integration.
Descubrimiento de datos	• Aunque no es muy fluida en sus características funcionales de descubrimiento de datos, permite mostrar líneas de tendencia y áreas subyacentes de datos. • Visor OLAP de arrastrar y soltar, con capacidades para representar información geográfica. Puede realizar cálculos al vuelo, editar fórmulas y otras tantas funcionalidades. Se pueden diseñar dashboards.
Colaboración	• La Suite es muy completa e incluye módulos para la publicación web y mobile de los cuadros de mando. • Es muy limitada en funcionalidades sociales y de colaboración, sin embargo es potente en aspectos de integración de las visualizaciones con otras plataformas como Liferay y Sharepoint.
Administración	• Aunque incluye herramientas poderosas de transformación, no es completamente intuitiva y delegable en un usuario funcional por lo que conserva una alta dependencia del área técnica. • Es open source y se distribuye bajo licencia GPL. No hay necesidad de acceder a una versión premium, dado que la Suit completa se ofrece de forma gratuita.

CATEGORÍA	CARACTERÍSTICA TÉCNICA	IMPORTANCIA	CUMPLIMIENTO	CALIFICACIÓN
Conexión a datos	Conexión a fuentes de datos del tipo DWH	8	3	24
	Conexión simultanea a varias fuentes de datos	7	1	7
Descubrimiento de datos	Metodos de exploración en profundidad (drill-down)	6	2	12
	Cuadros de mando avanzados con riqueza visual e interactivos	5	3	15
Colaboración	Publicación de dashboards en un Cloud BI	3	3	9
	Capacidadades sociales y de colaboración	2	2	4
Administración	Auto-servicio y preparación de datos	4	1	4
	Facilidad para escalar a un versión premium	1	3	3
			TOTAL	**9,8**

Tabla 14. Análisis y evaluación de la herramienta Pentaho Community Dashboard

2.7. Arquitectura elegida

Teniendo en cuenta la metodología de análisis y evaluación de herramientas y los criterios técnicos pensados en la escogencia de las tecnologías open source o gratuita más adaptable a las necesidades de un Sistema BI para una empresa mediana de marketing digital, los resultados nos permiten seleccionar las siguientes herramientas:

Capa de integración de datos	
HERRAMIENTA	CALIFICACIÓN
Pentaho Data Integration	11.1
Talend Open Studio	12.9
CoverETL Community Edition	5.6

Tabla 15. Resumen evaluación herramientas capa integración de datos

Talend Open Studio, destaca por ser la solución más completa de las evaluadas, que incorpora de forma nativa las principales funcionalidades para la diagramación, transformación, control y trazabilidad de procesos ETL, reduciendo los tiempos de adopción y uso.

Capa de bodega de datos	
HERRAMIENTA	CALIFICACIÓN
PostgresSQL	13.5
Oracle MySQL	7.6
MariaDB	8.4

Tabla 16. Resumen evaluación herramientas capa bodega de datos

PostgreSQL, cumple con todas las características para soportar el modelamiento y gestión de un Data Warehouse, en particular porque está diseñado para almacenar grandes volúmenes de datos y está optimizado para la concurrencia de lectura y escritura, lo que supone una convivencia entre el cambio evolutivo de la bodega de datos con nueva información y sus tareas prioritarias de consulta. Además, el hecho de poder crear nuevas tipologías de datos le da un plus a la hora de trabajar con dimensiones geográficas y complejas jerarquías de datos.

Capa de visualización de datos	
HERRAMIENTA	CALIFICACIÓN
Microsoft Power BI	12.5
Tableau Public	9.5
Pentaho Community Dashboard	9.8

Tabla 17. Resumen evaluación herramientas capa visualización de datos

Microsoft Power BI, está a la altura de las herramientas top para la analítica y visualización avanzada de datos, porque mezcla resultados gráficos muy estilizados con la potencia del descubrimiento profundo de datos.

Además, comprende bastante bien el concepto de la analítica de autoservicio, es decir, se adapta al modo de pensar de las personas. Power BI abstrae al usuario de conocimientos avanzados de bases de datos o programación, porque su base de usuarios normalmente proviene de un mundo ya conocido (Excel) por lo tanto recorren una curva de aprendizaje muy corta y producen nuevo conocimiento e información en menos tiempo.

Power BI es una herramienta ajustada a la dinámica del sector de la publicidad y el mercadeo, muy bien soportada por Microsoft que además brinda una oferta comercial muy competitiva por su versión premium.

3. Diseño e Implementación del Data Warehouse

Debido a las campañas de mercadeo y publicidad que han sido desplegadas en medios digitales como Facebook, Youtube, Twitter e Instagram, se requiere analizar el resultado de estas, y a partir de estos datos generar unos informes que le permitan al equipo de publicidad perfilar mejor al consumidor, identificando los rangos de edad, género, localización y hasta gustos que son más proclives a mejorar las tasas de repuestas positivas a un anuncio.

Para esto, se ha determinado diseñar un Data Warehouse, donde se almacenarán, de forma estructurada, los datos obtenidos que detallan el comportamiento y entregan información acerca de los usuarios que consumen las campañas.

En este capítulo se detallarán las etapas que conllevan la comprensión del caso y el posterior diseño del Data Warehouse, se describirán lo datos del análisis publicitario; se expondrá el diseño del modelo conceptual de datos, con el cual se identifican las vistas necesarias para dar respuesta a las preguntas; se expondrá el diseño del modelo lógico de los datos, con el cual se identificarán las métricas necesarias y, por último, se expondrá el modelo físico de los datos, que materializa el diseño en un gestor de base de datos.

3.1. Comprensión de los datos

Se ha proporcionado una base de datos en archivos Excel que contienen información sobre las localizaciones de los usuarios de las plataformas digitales donde se desplegaron las campañas. El tipo de información contenida en cada hoja se explicará a continuación:

Productos: Contiene los nombres de los productos anunciados dentro de la campaña publicitaria, relacionados con el nombre de la familia a la que pertenecen.

Zonas: Contiene el nombre de las zonas con su respectivo nombre de la ciudad y código postal, donde la campaña publicitaria tiene cobertura.

Facebook: Contiene datos del consolidado del número de visualizaciones (prints) y el número de clicks (hits) de los anuncios desplegados en la plataforma Facebook, los cuales han sido agrupados por fecha, código postal de la zona, producto, rango de edad, género/sexo y gustos.

Youtube: Contiene datos del consolidado del número de visualizaciones (prints) y el número de clicks (hits) de los anuncios desplegados en la plataforma Youtube, los cuales han sido agrupados por fecha, código postal de la zona, producto, rango de edad, género/sexo y gustos.

Instagram: Contiene datos del consolidado del número de visualizaciones (prints) y el número de clicks (hits) de los anuncios desplegados en la

plataforma Instagram, los cuales han sido agrupados por fecha, código postal de la zona, producto, rango de edad, género/sexo y gustos.

Twitter: Contiene datos del consolidado del número de visualizaciones (prints) y el número de clicks (hits) de los anuncios desplegados en la plataforma Twitter, los cuales han sido agrupados por fecha, código postal de la zona, producto, rango de edad, género/sexo y gustos.

3.2. Modelo conceptual de los datos

El modelamiento conceptual es el primer paso para el diseño del Data Warehouse, donde se identificarán las tablas de hechos y las dimensiones necesarias para analizar los datos desde el punto de vista de los analistas del negocio, que en el caso de estudio corresponde al equipo de trabajo de marketing digital.

El método aplicado estará basado en el modelo dimensional que se describió por primera vez en el año 1996 por Ralph Kimball como propuesta para el diseño de un Data Warehouse partiendo de la visión multidimensional que los usuarios tienen de los datos empresariales cuando se enfrentan a ellos con propósitos de análisis.

El análisis multidimensional consiste en analizar datos que hacen referencia a hechos, desde la perspectiva de sus componentes o dimensiones utilizando para para ello algún tipo de métrica o medida de negocio” [10].

Los “hechos” es la representación en el Data Warehouse de los procesos del negocio de la organización, se reconocen porque siempre tienen asociada una fecha y no se modifican ni se eliminan para no perder la historia. En términos prácticos y en relación con el caso de estudio, el hecho será el “anuncio” (campaña publicitaria). Adicional a la definición del hecho, se deberán identificar las “métricas” del hecho, que serán los indicadores de negocio del proceso a modelar, es decir, los conceptos cuantificables para medir al proceso de negocio [11]. En el caso de estudio, las métricas serán: el número de visualizaciones (prints) y el número de clicks (hints) de un anuncio desplegado en una plataforma digital específica.

De forma complementaria, y como una “métrica derivada”, se agregará el indicador CTR, o tasa de clicks del anuncio, el cual se obtiene como un cálculo, dividendo el número de visualizaciones totales por el total de clicks recibidos.

Luego se identifican las “dimensiones”, que dentro del diseño del Data Warehouse corresponde al punto de vista desde el cual se puede analizar el hecho de cierto proceso de negocio [12]. En nuestro caso de estudio, las campañas publicitarias (anuncios), como hecho, pueden ser analizadas desde el punto de vista del producto anunciado, la ciudad del usuario consumidor del producto, el día en el que tuvo actividad la campaña publicitaria y en que plataforma digital, por lo tanto, producto, ciudad, tiempo y plataforma serán las dimensiones.

Este modelamiento dimensional permitirá validar el proceso de negocio, es decir, que la publicidad del producto basada en el indicador CTR es favorable dependiendo del punto de vista, también llamado vista, desde el cual se analiza. Soportado en lo anterior, podríamos considerar que el rango de edad, el género y los gustos del usuario al que se proyecta un anuncio, son también un punto de vista a considerar, por lo cual los agruparemos en una dimensión a la que llamaremos perfil.

Ahora que ya tenemos identificado el hecho, sus métricas, sus dimensiones y algunos atributos que ayudaran a su analítica, es hora de diagramarlo como un esquema dimensional, donde se puede adoptar la forma de "estrella" o de "copo de nieve", este último también conocido como "snow blake". Para el caso de estudio, partiremos inicialmente de un esquema en estrella, como se muestra a continuación, y determinaremos más adelante si requerimos evolucionar a un esquema tipo copo de nieve.

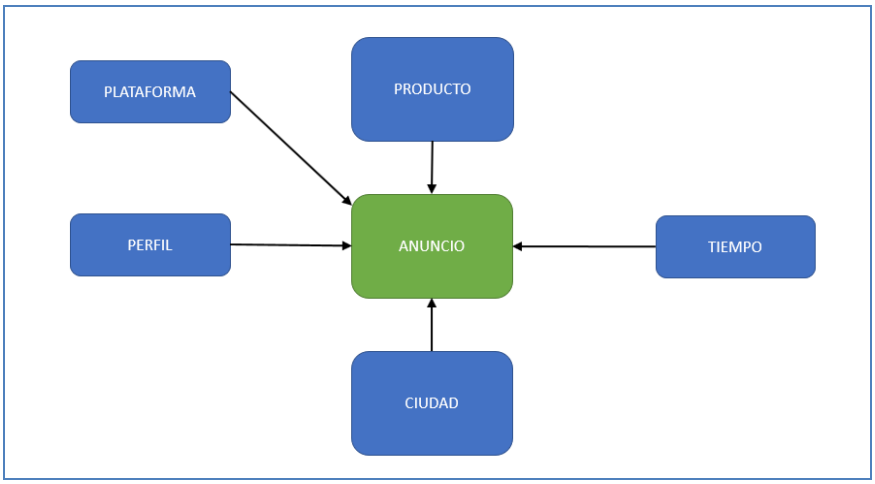


Figura 4. Modelo conceptual en estrella

Podemos ver como en la parte central tenemos la tabla de hechos, y el resto de las tablas que la rodean son las que se conocen como tablas de dimensiones.

El esquema "copo de nieve", se emplea cuando se decide normalizar las tablas de dimensiones para eliminar campos redundantes o armar jerarquías [12].

Para nuestro caso de estudio, podríamos tomar la decisión de normalizar la tabla producto, abstrayendo la "familia" relacionada al producto; al igual que la dimensión ciudad, abstrayendo la "zona", por lo cual el modelo conceptual originalmente en estrella evolucionaría a un modelo del tipo copo de nieve, como se muestra a continuación:

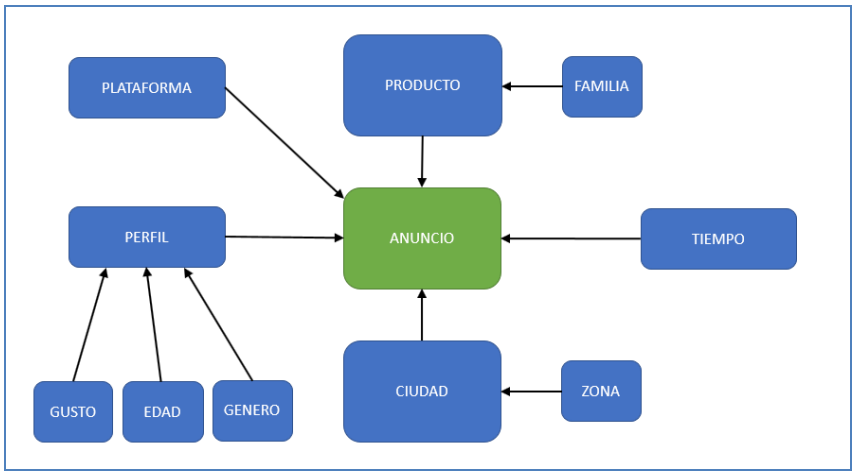


Figura 5 Modelo conceptual en copo de nieve

Normalizar una tabla de dimensión, es una decisión de diseño, cuyo propósito es dar respuesta rápida a las consultas de los usuarios del sistema BI, no propiamente ahorrar espacio.

Generalmente el espacio que se logra ahorrar cuando se normaliza la dimensión es insignificante al volumen de registros almacenados en la tabla de hechos, mientras el impacto en los tiempos de respuesta es apreciable.

3.3. Diseño del modelo lógico de los datos

Luego de realizar el modelo conceptual de los datos se pasa a diseñar el modelo lógico de los datos, donde se indican las métricas de las tablas de hechos, sus demás atributos y dimensiones, así como los atributos de estas dimensiones.

La tabla de hechos tiene su llave primaria para identificar cada registro de manera única, además de las diferentes llaves foráneas que la relaciona con las diferentes tablas de dimensión, así como las métricas para medir las visualizaciones y clicks en las diferentes plataformas digitales.

A continuación, la definición lógica de la tabla de hechos, que consolidará la información de las campañas realizadas en las diferentes plataformas digitales:

HCH_ANUNCIO: Tabla de hechos con el número de visualizaciones y clicks que se realizaron por día en los anuncios de productos por parte de ciertos perfiles de usuarios ubicados en determinadas zonas.

TABLA DE HECHO	LLAVE PRIMARIA	LLAVE FORANEA	MÉTRICA
hch_anuncio	id_anuncio	id_tiempo	prints
		id_plataforma	hits
		id_perfil	ctr
		id_ciudad	
		id_producto	

Tabla 18. Tabla de Hecho HCH_ANUNCIO

A continuación, los atributos de cada una de las tablas de dimensión:

DIM_PRODUCTO: Tabla de dimensión con los productos anunciados en las campañas publicitarías y la familia/categoría a las que pertenecen.

TABLA DE DIMENSIÓN	LLAVE PRIMARIA	ATRIBUTO	
dim_producto	id_producto	nombre_producto	niveles de jerarquía
		familia_producto	

Tabla 19. Tabla de Dimensión DIM_PRODUCTO

DIM_CIUDAD: Tabla de dimensión con las ciudades donde se anuncian las campañas publicitarias y la zona y código postal a la que pertenece.

TABLA DE DIMENSIÓN	LLAVE PRIMARIA	ATRIBUTO	
dim_ciudad	id_ciudad	nombre_ciudad	niveles de jerarquía
		zona	
		codigo_postal	rol geográfico

Tabla 20. Tabla de Dimensión DIM_CIUDAD

DIM_TIEMPO: Tabla de dimensión con las fechas en las que se anuncian las campañas publicitarias, con desagregación por día, semana, mes, trimestre y año.

TABLA DE DIMENSIÓN	LLAVE PRIMARIA	ATRIBUTO	
dim_tiempo	id_tiempo	dia	niveles de jerarquía
		semana	
		mes	
		trimestre	
		año	

Tabla 21. Tabla de Dimensión DIM_TIEMPO

DIM_PERFIL: Tabla de dimensión con el perfil de usuarios que interactúan con los anuncios de campañas publicitarias, teniendo en cuenta aspectos como el gusto/interés, el rango de edad y el género/sexo.

TABLA DE DIMENSIÓN	LLAVE PRIMARIA	ATRIBUTO	
dim_perfil	id_perfil	edad	niveles de jerarquía
		genero	
		gusto	

Tabla 22. Tabla de Dimensión DIM_PERFIL

DIM_PLATAFORMA: Tabla de dimensión con la plataforma digital donde se despliega el anuncio de la campaña publicitaria.

TABLA DE DIMENSIÓN	LLAVE PRIMARIA	ATRIBUTO
dim_plataforma	id_plataforma	nombre_plataforma

Tabla 23. Tabla de Dimensión DIM_PLATAFORMA

A partir de la tabla de hechos y las tablas de dimensión, se obtiene el siguiente modelo físico de los datos:

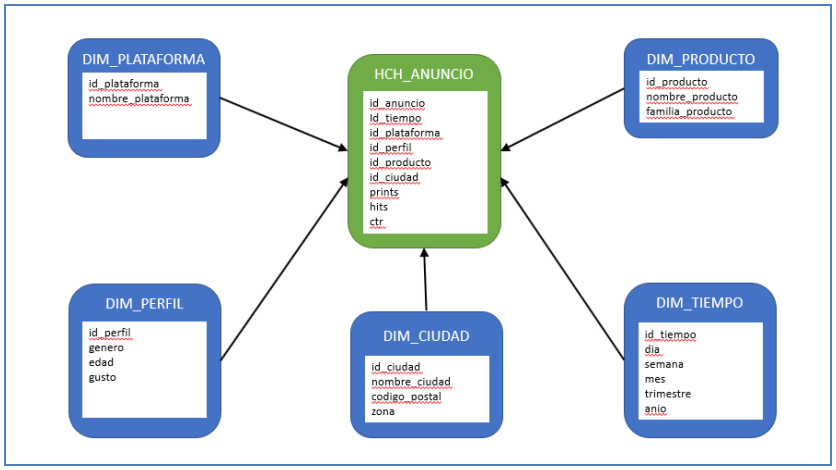


Figura 6. Modelo lógico de datos

En un ámbito más amplio al que hemos definido dentro del alcance del diseño de este Data Warehouse, esta tabla de hechos con sus dimensiones relacionadas se podría comprender como el “Data Mart Campañas Publicitarias”. Este Data Mart, perteneciente al Data Warehouse, se poblará con datos provenientes de los sistemas operacionales de cada una de las plataformas digitales mediante el uso de procesos ETL, para su posterior explotación a través de herramientas de inteligencia de negocios y podría hacer parte de un esquema más grande si al Data Warehouse lo complementamos con otros Data Marts adicionales de propósito específico, como podría serlo el “Data Mart Ventas” o el “Data Mart Clientes”, que podrían maximizar las oportunidades para realizar análisis y tomar decisiones.

Cada uno de estos DATA MARTS tendrá a su vez que estar conformados con sus correspondientes tablas de hechos, métricas e indicadores, y relacionados a su vez con sus propias dimensiones.

3.4. Diseño del modelo físico de los datos

Una vez realizado el modelamiento lógico de los datos, el siguiente y último paso para crear el Data Warehouse es la creación del modelo físico de datos.

Para llevar a cabo este paso y teniendo en cuenta que usaremos PostgreSQL como motor relacional de base de datos, se trabajará con su herramienta de diagramación visual para PostgreSQL llamada **pgModeler** en su versión 0.9.0.

La creación del modelo físico estará basada en el diseño del modelo lógico obtenido anteriormente, con ello se define cada formato de las llaves y atributos de las tablas. A continuación, se muestra el diseño del modelo físico:

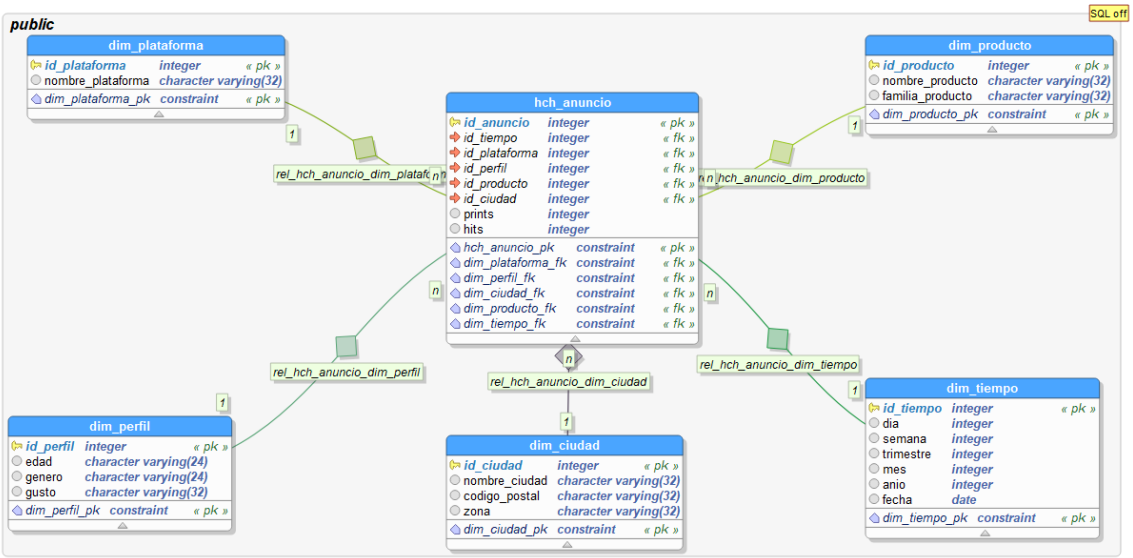


Figura 7. Modelo físico de datos

3.5. Implementación del Data Warehouse

Una vez realizado el modelo físico de datos, tenemos el insumo fundamental a nivel de estructura de datos para implementar el Data Warehouse, el cual se implementará como un schema de base de datos dentro del motor de base de datos PostgreSQL versión 10, al que llamaremos ***dwh***. El primer paso será crear la base de datos ***dwh*** en PostgreSQL como un (1) schema vacío sin tablas ni demás objetos, para esto utilizaremos la herramienta de administración pgAdmin 4 versión 3.0:

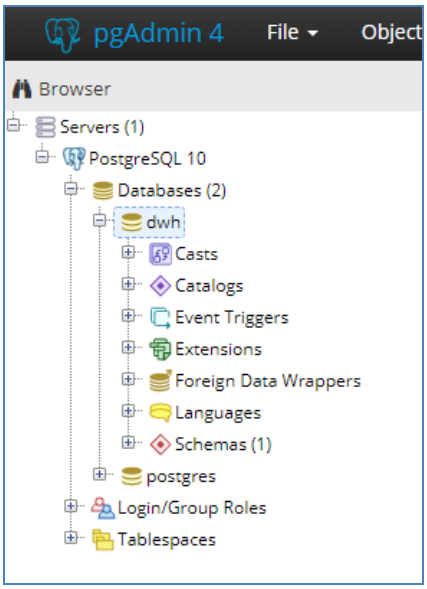


Figura 8. Base de datos dwh

Desde el modelador *pgModeler*, crearemos una conexión a esta base de datos, a través de la opción *Settings* → *Connections* → *New* → *General*, y se indicaran los parámetros básicos de la conexión:

- Nombre de la conexión definida en pgModeler (*Connection Alias*): dwh
- Nombre de la base de datos (*Connection DB*): dwh
- Puerto de conexión (*Host/Port*): 5432
- Servidor de base de datos (*Host/Port*): localhost
- Usuario propietario de la base de datos (*User*): postgres
- Contraseña del usuario propietario (*Password*): *********

Y por último damos clic en el botón *Apply*. Como se muestra en la siguiente imagen:

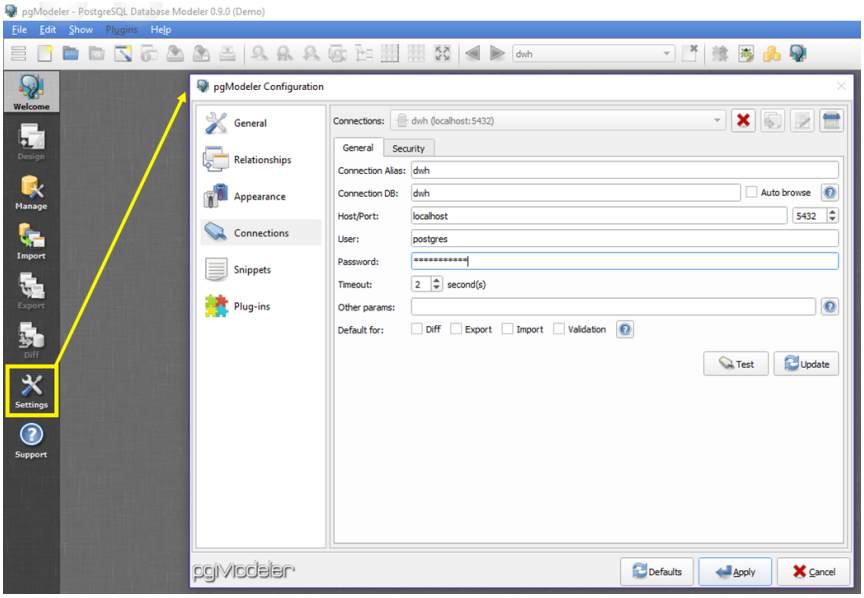


Figura 9. Conexión Base de Datos

Para crear las tablas y demás objetos, desde *pgModeler* realizamos la exportación del modelo a través de la opción *Export* en cuya ventana emergente establecemos la conexión a la base de datos y damos clic en el botón Export:

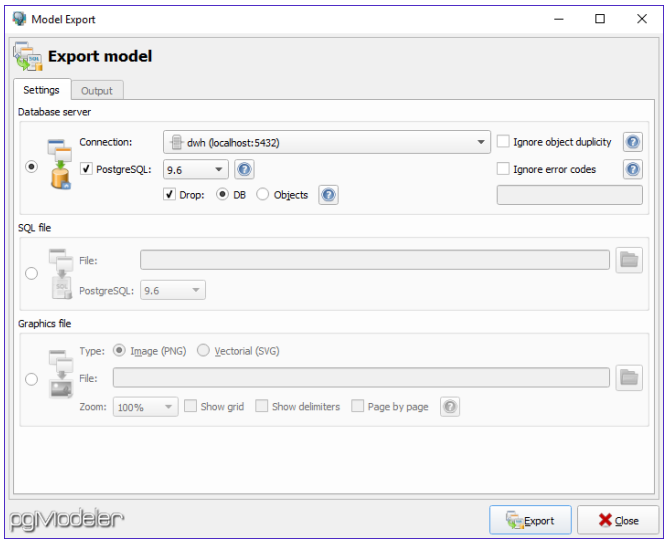


Figura 10. Exportación del modelo

Al final del proceso, se nos indicará la creación de todos los objetos en la base de datos: tablas, llaves primarias, llaves foráneas, etc, los cuales podemos verificar a través de la herramienta de administración de PostgreSQL *pgAdmin 4*:

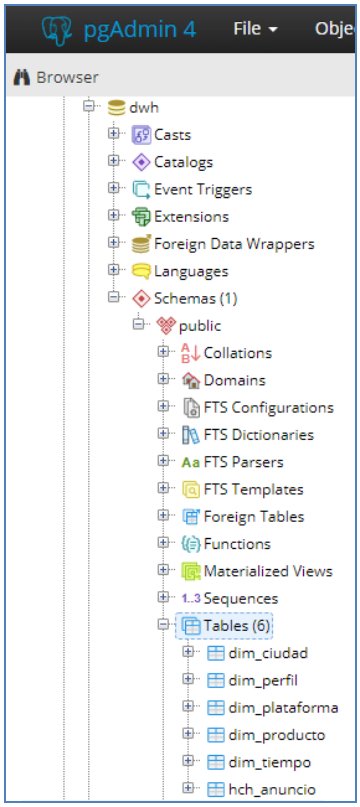


Figura 11. Objetos de la Base de datos

Las tablas se crearán en PostgreSQL vacías, sin datos, será responsabilidad de los procesos ETL alimentarlas con la información correcta a partir de las fuentes de datos proporcionadas.

4. Diseño e implementación de los procesos ETL

Una vez que se diseña e implementa la estructura de la bodega de datos (Data Warehouse), el siguiente paso es el diseño e implementación de los procesos ETL para la extracción, transformación y carga de los datos dentro del Data Warehouse.

El proceso ETL toma los datos de los diferentes orígenes de información y fuentes de datos, luego se procesan y si es del caso se transforman, para finalmente ser cargados en las bodegas de información (Data Warehouse) que serán destinadas para el posterior análisis y visualización de los datos.

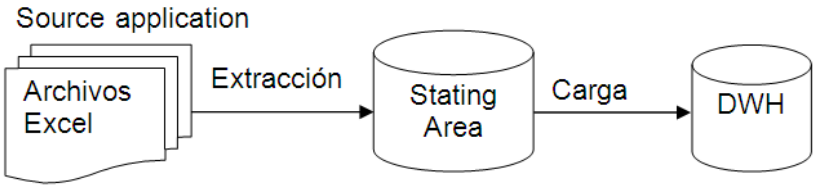


Figura 12. Objetos de la Base de datos

La herramienta que se usará para la integración de los datos es Talend Open Studio versión 7.0, que proporciona un entorno para modelar flujos ETL y un área de trabajo en el "heap memory" de una máquina virtual JAVA que se podrá acondicionar para procesar y transformar los datos que serán posteriormente cargados al Data Warehouse.

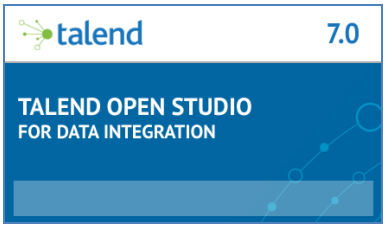


Figura 13. Talend Open Studio Data Integration, herramienta open source.

Lo primero que haremos con Talend es crear nuestro proyecto al que llamaremos *MasterUOC*.

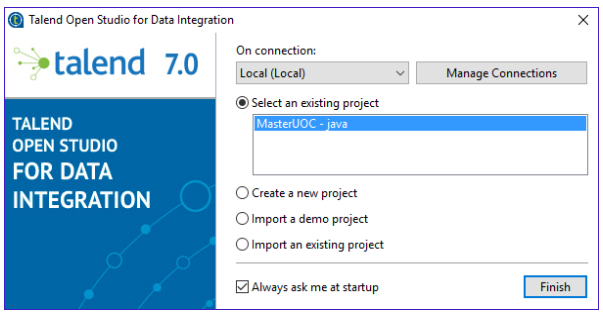


Figura 14. Creación del proyecto en Talend Open Studio

Después crearemos las conexiones y se definirán los esquemas de metadato de las fuentes en Excel desde las cuales se extraerán los datos:

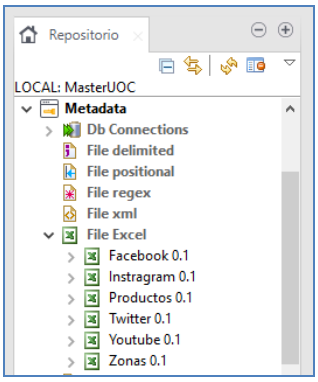


Figura 15. Conexiones a fuentes Excel en Talend Open Studio

También crearemos las conexiones y se definirán los esquemas de metadato de las tablas de dimensiones y la tabla de hechos en PostgreSQL, donde se cargarán los datos procesados y que en algunos casos también serán consultadas en algunos pasos del proceso ETL:

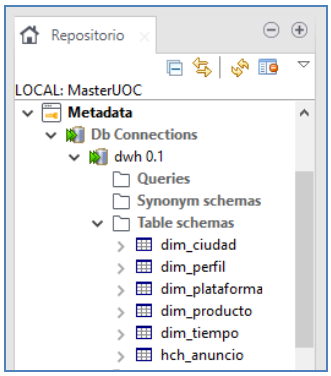


Figura 16. Conexiones a DWH de PostgreSQL en Talend Open Studio

En la sección ***Job Designs*** de Talend se crearán y modelaran los procesos ETL:

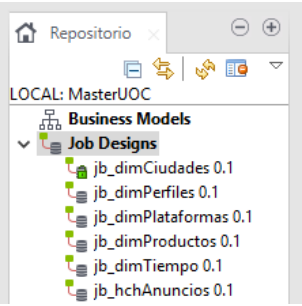


Figura 17. Listado de Jobs en Talend Open Studio, con el modelamiento de los flujos ETL.

Los Jobs deberán ser ejecutados bajo el siguiente orden:

- 1º: jb_dimCiudades
- 2º: jb_dimProductos
- 3º: jb_dimTiempo
- 4º: jb_dimPlataformas
- 5º: jb_dimPerfiles
- 6º: jb_hchAnuncios

4.1. Extracción, transformación y carga de la dimensión ciudad

Para cargar los datos de las ciudades se modeló en *Talend Open Studio* un proceso ETL llamado ***jb_dimCiudades***. El job extrae los datos a partir de un archivo Excel (*Zonas.xlsx*) y los carga, sin mayores transformaciones, en la tabla de dimensiones *dim_ciudad* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

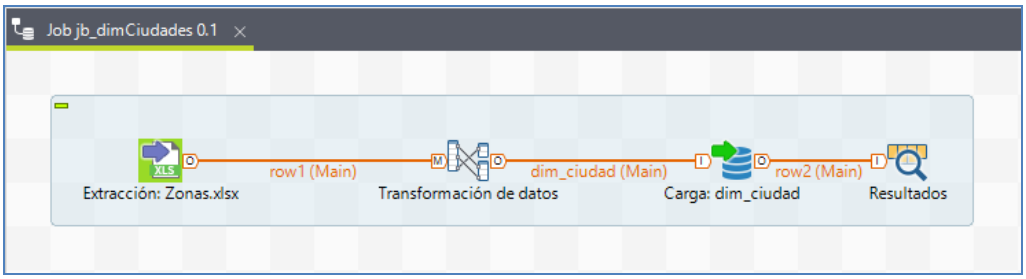


Figura 18. job jb_dimCiudades implementado en Talend Open Studio.

El proceso inicia con un componente ***tFileInputExcel*** donde se abre una conexión al archivo de Excel *Zonas.xlsx* y se extrae los datos de ciudad, zona y código postal:

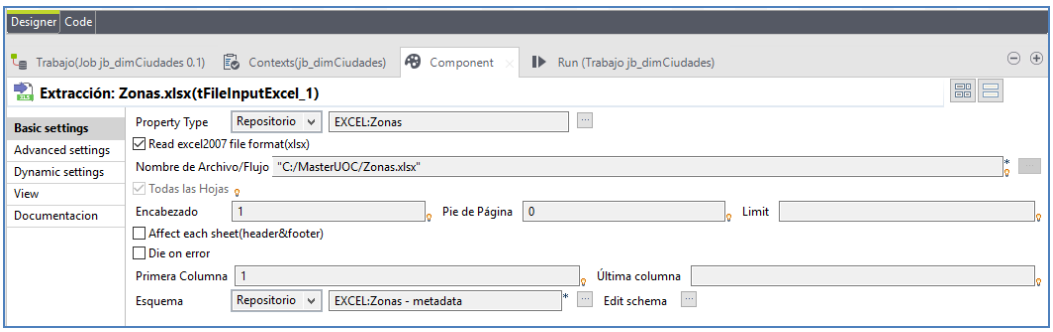


Figura 19. Componente tFileInputExcel para extraer los datos desde Zonas.xlsx

La salida es enviada a un componente ***tMap*** que se encargará de realizar el mapeo de campos entre la fuente y el destino (tabla de dimensión *dim_ciudad*), además de realizar una pequeña transformación en el campo *id_ciudad* donde se registrará una secuencia numérica:

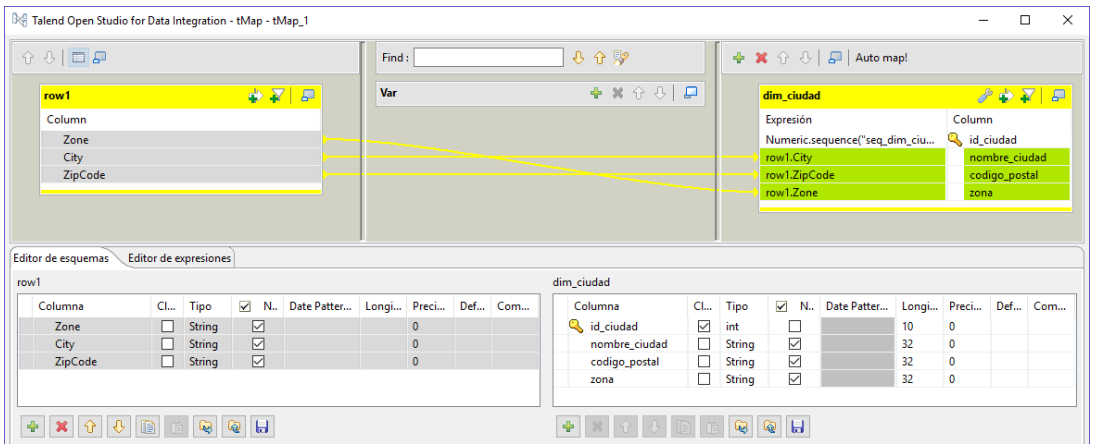


Figura 20. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla *dim_ciudad* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

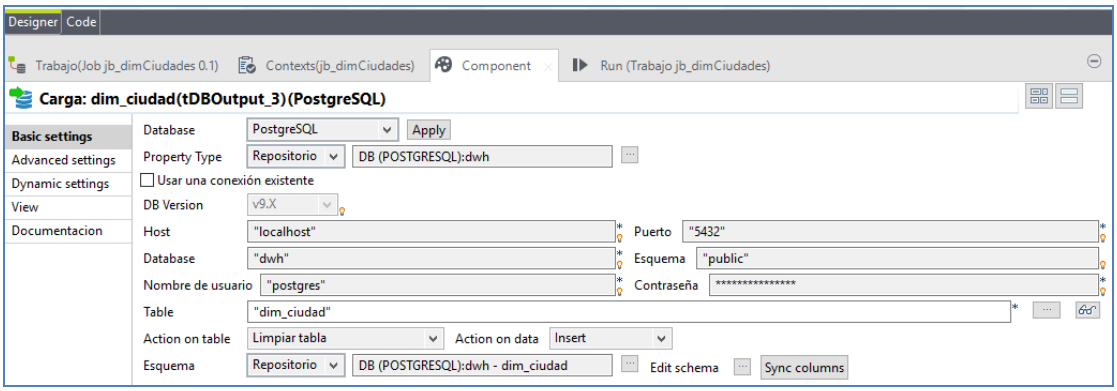


Figura 21. Componente tDBOutput para el cargue de datos en la tabla dim_ciudad.

El proceso ETL se ejecuta dando clic en el botón ***Run***:

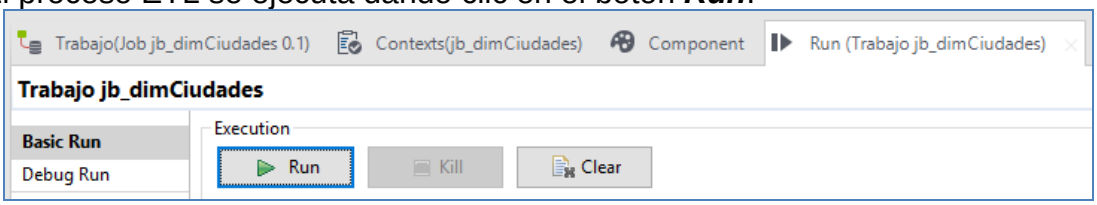


Figura 22. Ejecución proceso ETL jb_dimCiudades.

4.2. Extracción, transformación y carga de la dimensión producto

Para cargar los datos de los productos se modeló en *Talend Open Studio* un proceso ETL llamado ***jb_dimProductos***. El job extrae los datos a partir de un archivo Excel (*Productos.xlsx*) y los carga, sin mayores transformaciones, en la tabla de dimensiones *dim_producto* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

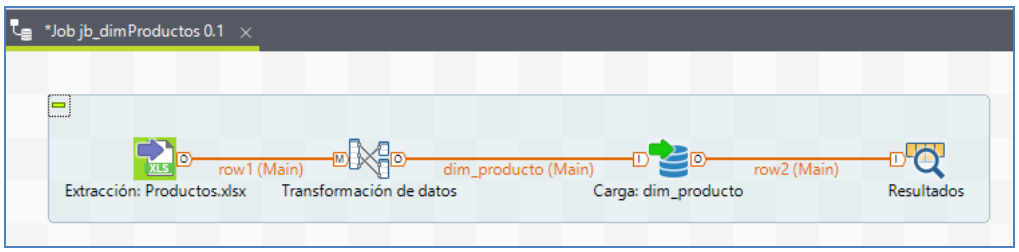


Figura 23. job jb_dimProductos implementado en Talend Open Studio.

El proceso inicia con un componente ***tFileInputExcel*** donde se abre una conexión al archivo de Excel *Productos.xlsx* y se extrae los datos de los productos y la familia del producto:

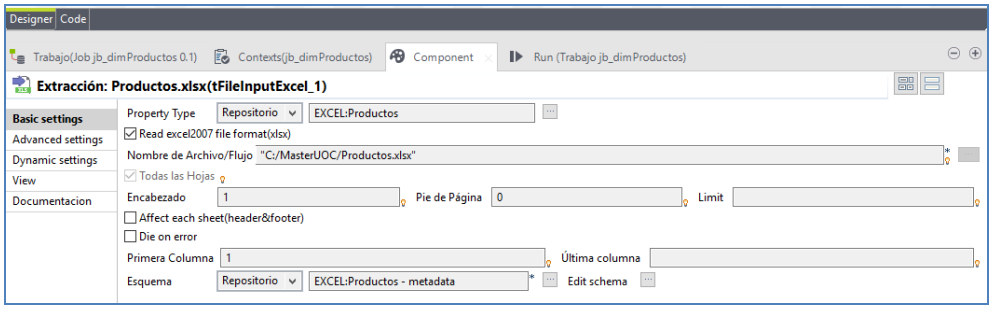


Figura 24. Componente tFileInputExcel para extraer los datos desde Productos.xlsx

La salida es enviada a un componente ***tMap*** que se encargará de realizar el mapeo de campos entre la fuente y el destino (tabla de dimensión *dim_producto*), además de realizar una pequeña transformación en el campo *id_producto* donde se registrará una secuencia numérica:

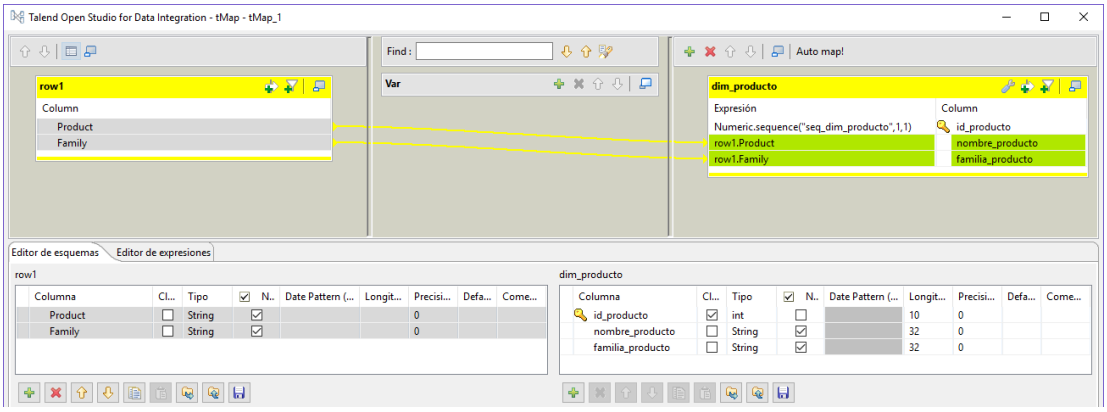


Figura 25. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla *dim_producto* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

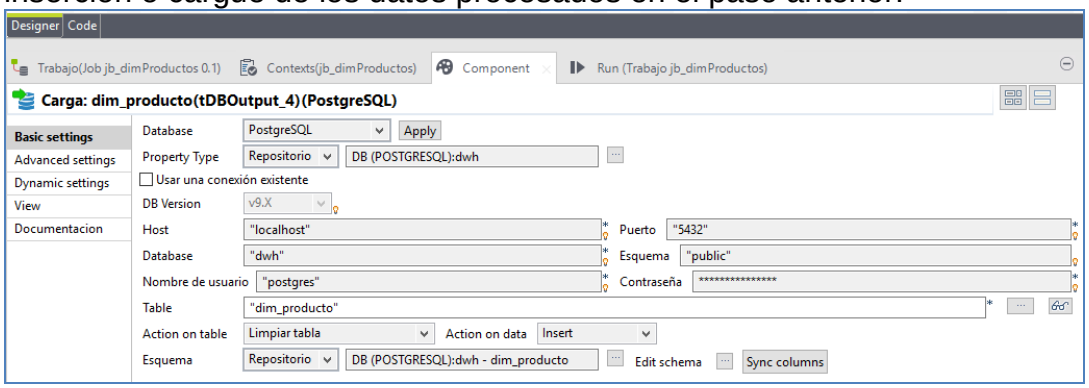


Figura 26. Componente tDBOutput para el cargue de datos en la tabla dim_producto.

El proceso ETL se ejecuta dando clic en el botón ***Run***:

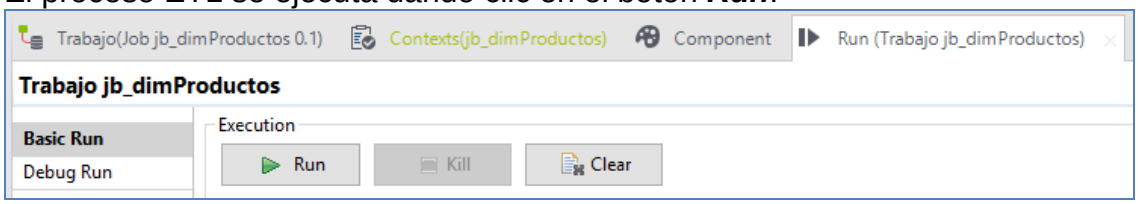


Figura 27. Ejecución proceso ETL jb_dimProductos.

4.3. Extracción, transformación y carga de la dimensión tiempo

Para cargar los datos de las fechas se modeló en *Talend Open Studio* un proceso ETL llamado ***jb_dimTiempo***. El job consolida las fechas con las que trabajará el data warehouse a partir de las fechas obtenidas de las mismas fuentes de datos proporcionadas por las plataformas digitales en archivos de Excel, realizando operaciones de transformación alrededor de la fecha para obtener las diferentes jerarquías de tiempo, previo al cargue de los datos transformados en la tabla de dimensiones *dim_tiempo* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

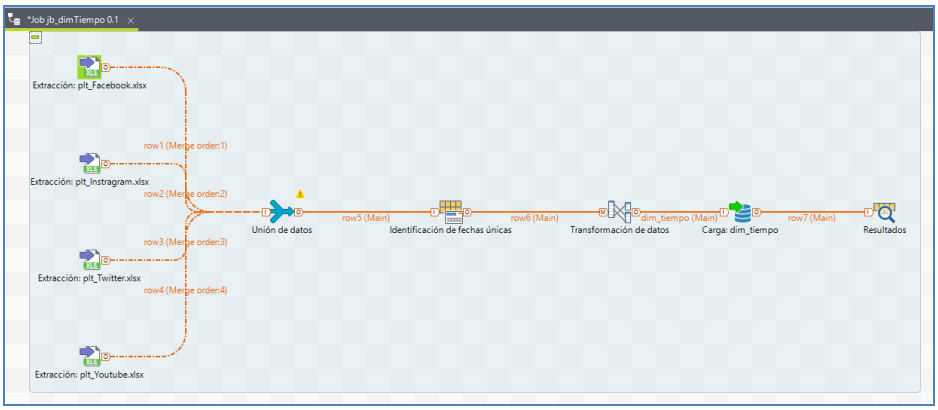


Figura 28. job jb_dimTiempo implementado en Talend Open Studio.

El proceso inicia con cuatro (4) componentes ***tFileInputExcel*** que en paralelo abren una conexión a los archivos de Excel *plt_Instagram.xlsx, plt_Facebook.xlsx, plt_Youtube.xlsx y plt_Twitter.xlsx* y se extraen sus contenidos:

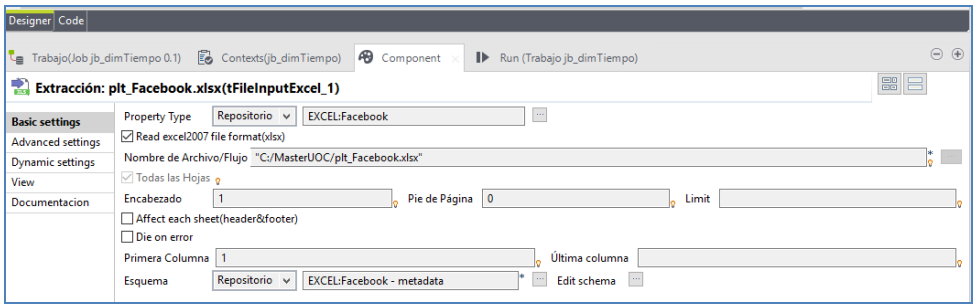


Figura 29. Componente tFileInputExcel para extraer los datos desde plt_Facebook.xlsx

La salida es enviada a un componente ***tUnit*** que mezcla los datos proporcionados por las diferentes fuentes de las plataformas digitales:

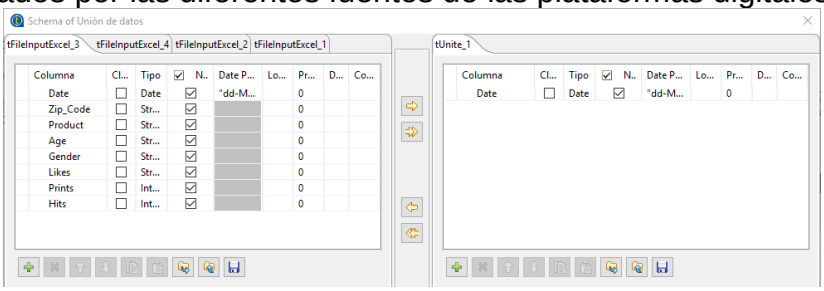


Figura 30. Componente tUnit para unir el contenido de las fuentes de las plataformas digitales.

La salida es enviada a un componente ***tAggregateRow*** que agrupa los datos por los valores del campo Date, eliminando las filas duplicadas y consolidando un listado de fechas únicas:

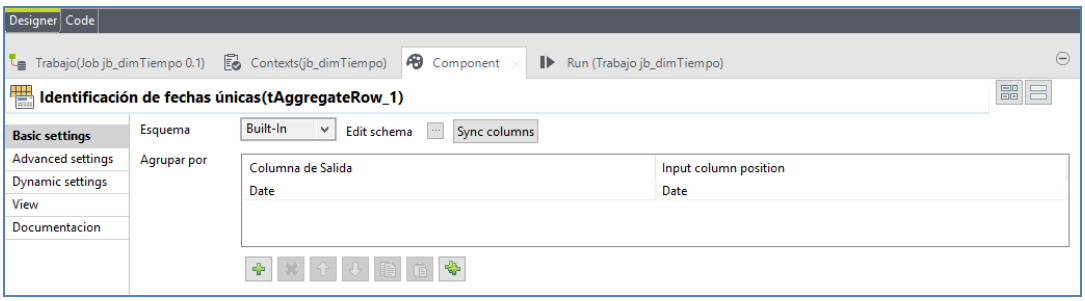


Figura 31. Componente tAggregateRow para agrupar los datos por el campo Date.

La salida es enviada a un componente ***tMap*** que se encargará de realizar el mapeo de campos entre la fuente y el destino (tabla de dimensión *dim_tiempo*). Para esto realizará una pequeña transformación en el campo *id_tiempo* donde se registrará una secuencia numérica, y soportado en el campo Date realizará transformaciones usando funciones de transformación de datos tipo fecha para obtener los valores de los diferentes niveles de jerarquía: día, mes, trimestre, año y semana del mes:

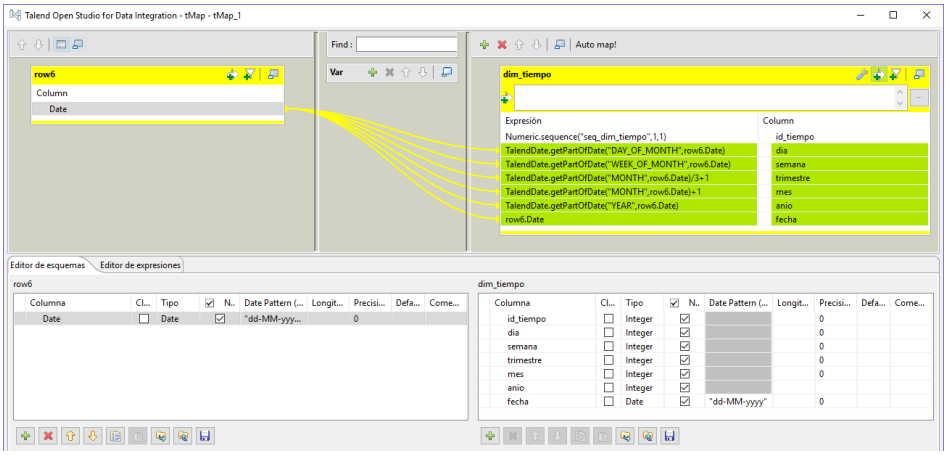


Figura 32. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla *dim_tiempo* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

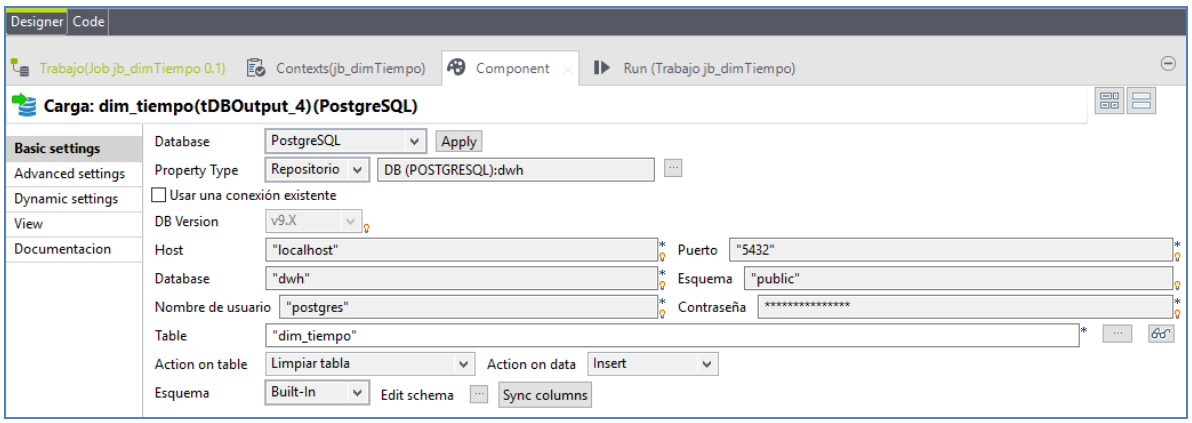


Figura 33. Componente tDBOutput para el cargue de datos en la tabla dim_tiempo.

El proceso ETL se ejecuta dando clic en el botón ***Run***:

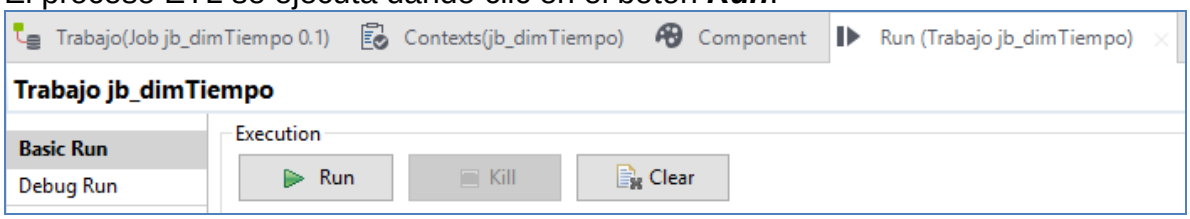


Figura 34. Ejecución proceso ETL jb_dimTiempo.

4.4. Extracción, transformación y carga de la dimensión plataforma

Para cargar los datos de las plataformas digitales donde se desplegó publicidad se modeló en *Talend Open Studio* un proceso ETL llamado ***jb_dimPlataformas***. El job extrae los datos a partir de las fuentes en Excel

(*plt_Facebook.xlsx, plt_Instagram, plt_Twitter y plt_Youtube*) y los carga en la tabla de dimensiones *dim_plataforma* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

Figura 35. job jb_dimPlataformas implementado en Talend Open Studio.

El proceso inicia con un componente ***tFileList*** que a partir de una carpeta de archivos obtiene el listado de aquellos que cumplan con el patrón de un archivo Excel proporcionado por una plataforma: comienzan con el prefijo *"plt_"* y terminan en una extensión *".xlsx"*:

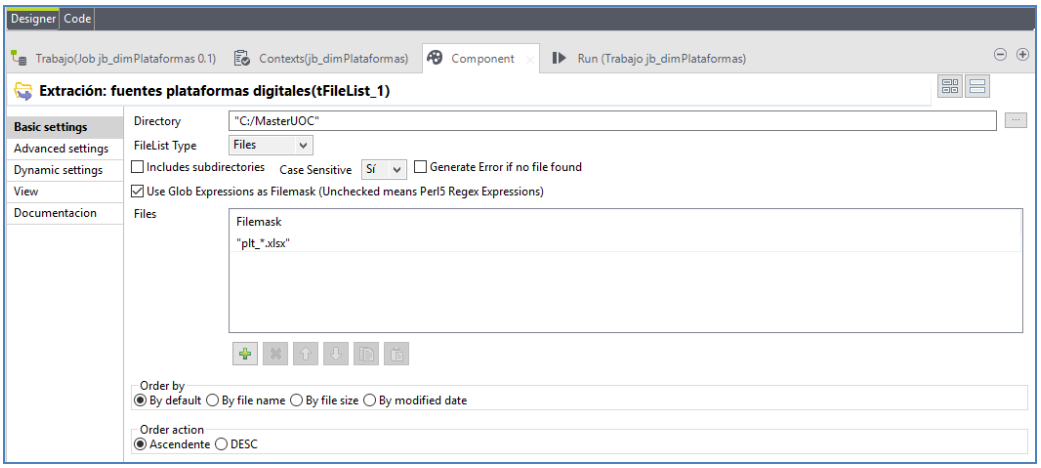


Figura 36. Componente tFileList para listar los archivos plt_[NOMBRE-PLATAFORMA].xlsx

En cada iteración, se obtendrá un archivo de una plataforma digital, el cual será abierto a través del componente ***tFileInputExcel*** y de su contenido se obtendrá el campo ***Plataforma*** que tendrá el descriptor o nombre de la plataforma digital.

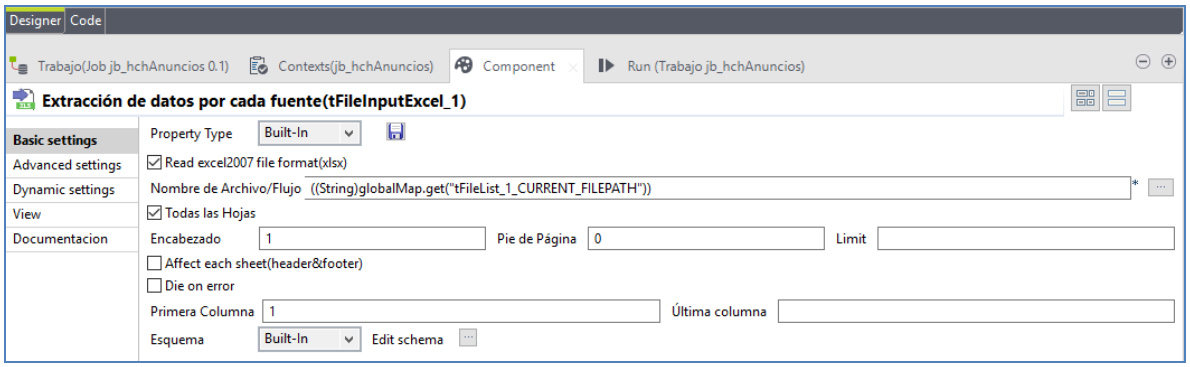


Figura 37. Extracción de los datos de las fuentes proporcionadas por las plataformas digitales.

La salida de cada iteración se unirá a través del componente ***tUnit*** y después se agrupará por nombre de plataforma digital a través del componente ***tAggregateRow***, eliminando duplicados y conservando los sólo el listado de

nombres únicos de plataformas digitales, que serán después cargados en la tabla de dimensiones respectiva.

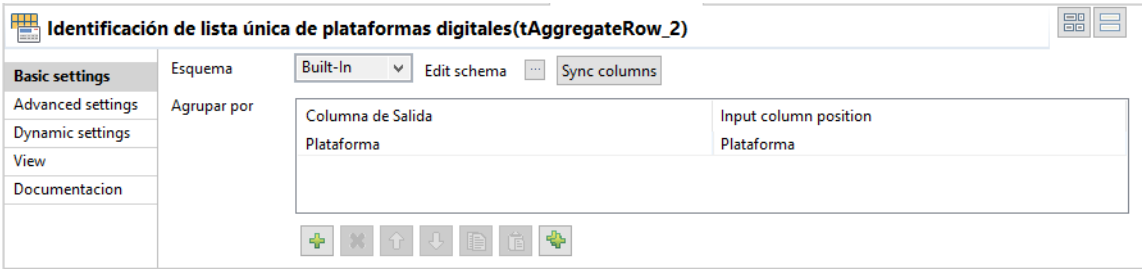


Figura 38. Componente tAggregateRow

Esa variable es enviada como entrada al componente ***tMap*** que se encargará de realizar el mapeo de ese campo al campo destino en la tabla de dimensión *dim_plataforma*, además de realizar una pequeña transformación en el campo *id_plataforma* donde se registrará una secuencia numérica:

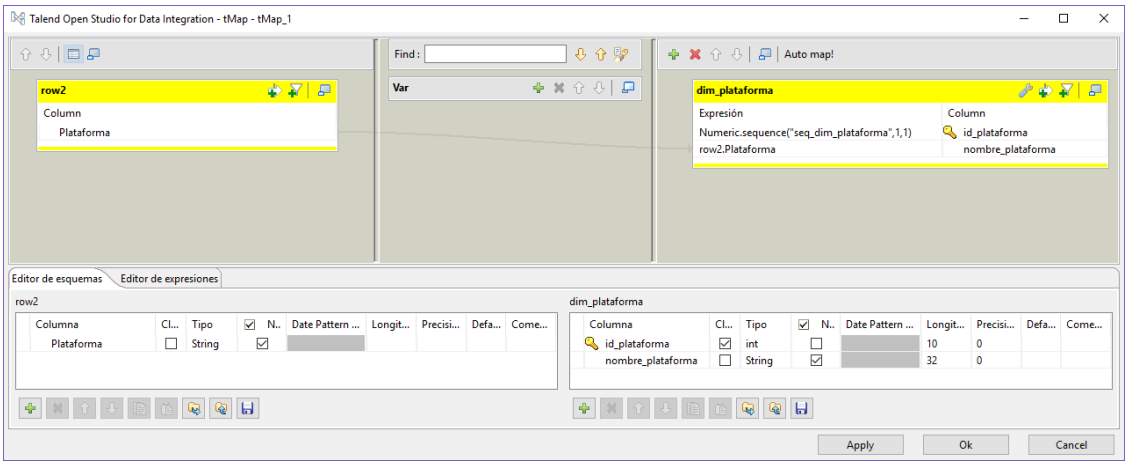


Figura 39. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla *dim_plataforma* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

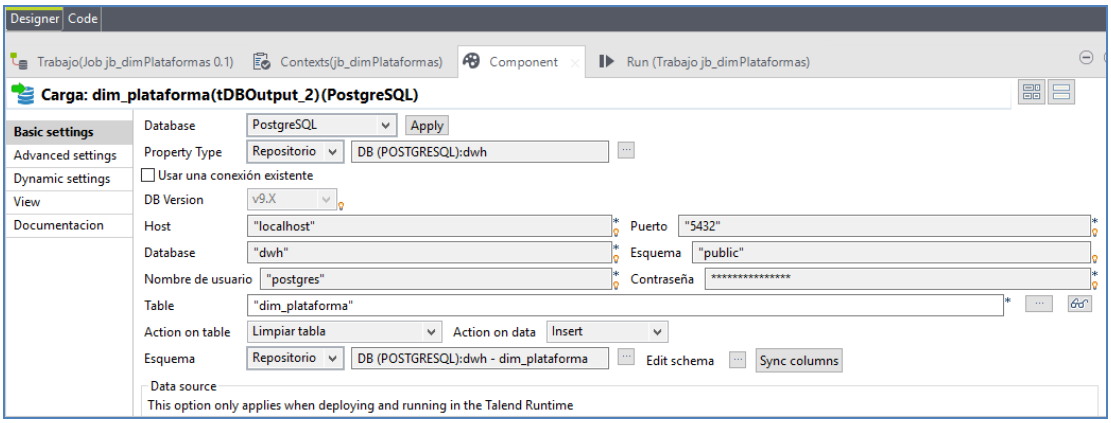


Figura 40. Componente tDBOutput para el cargue de datos en la tabla dim_plataforma.

El proceso ETL se ejecuta dando clic en el botón ***Run***:

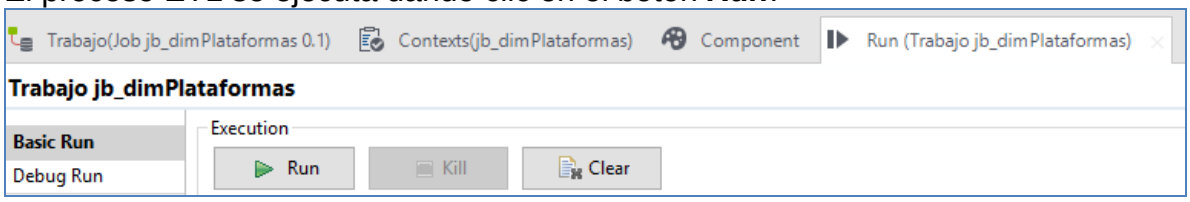


Figura 41. Ejecución proceso ETL jb_dimPlataformas.

4.5. Extracción, transformación y carga de la dimensión perfil

Para cargar los datos de las ciudades se modeló en *Talend Open Studio* un proceso ETL llamado ***jb_dimPerfiles***. El job consolida los datos de rango de edad, género/sexo y gusto/afición de los usuarios que respondieron a la campaña de marketing digital. Estos datos juntos componen el perfil del usuario. La tabla de dimensiones Perfil que compone el data warehouse se armará a partir de estos tres datos (edad, género y gusto) obtenidos de las mismas fuentes de datos proporcionadas por las plataformas digitales en archivos de Excel, realizando operaciones de transformación alrededor de los tres campos ya mencionados, previo al cargue de los datos transformados en la tabla de dimensiones *dim_perfil* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

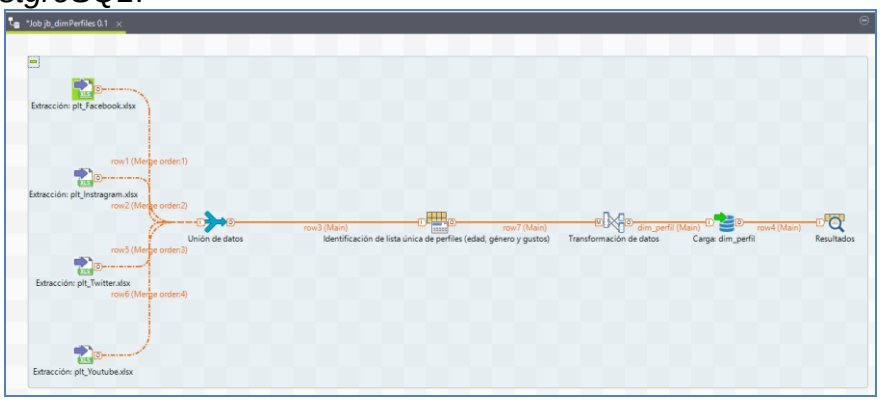


Figura 42. job jb_dimPerfiles implementado en Talend Open Studio.

El proceso inicia con cuatro (4) componentes ***tFileInputExcel*** que en paralelo abren una conexión a los archivos de Excel *plt_Instagram.xlsx, plt_Facebook.xlsx, plt_Youtube.xlsx y plt_Twitter.xlsx* y se extraen sus contenidos:

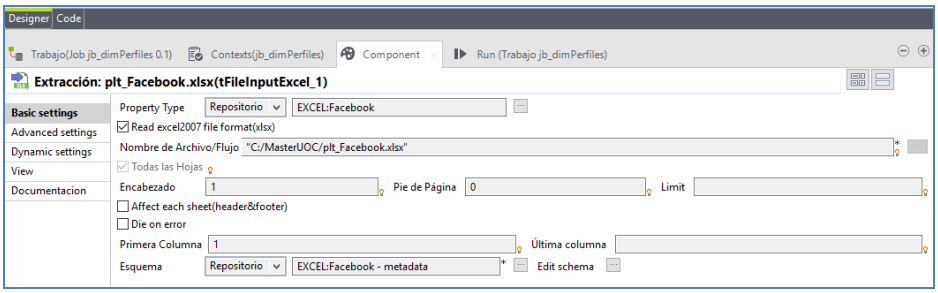


Figura 43. Componente tFileInputExcel para extraer los datos desde plt_Facebook.xlsx

La salida es enviada a un componente ***tUnit*** que mezcla los datos proporcionados por las diferentes fuentes de las plataformas digitales:

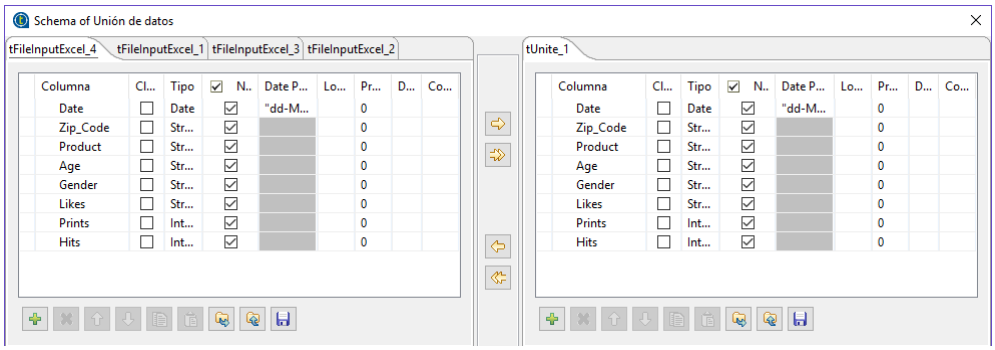


Figura 44. Componente tUnit para unir el contenido de las fuentes de las plataformas digitales.

La salida es enviada a un componente ***tAggregateRow*** que agrupa los datos por los valores de los campos *Edad, Género y Gusto*, eliminando las filas duplicadas y consolidando un listado de perfiles únicos:

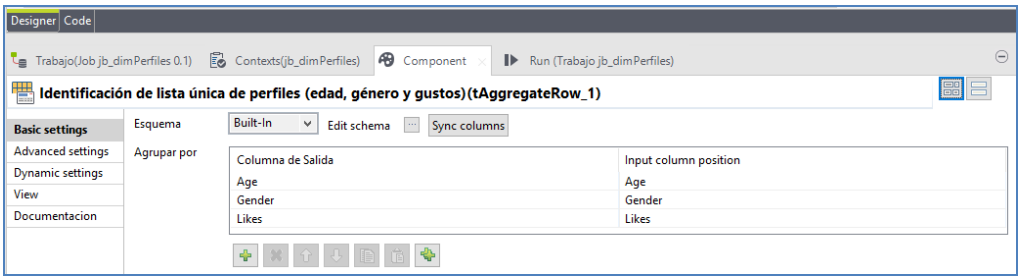


Figura 45. Componente tAggregateRow para agrupar los datos por los tres campos del perfil.

La salida es enviada a un componente ***tMap*** que se encargará de realizar el mapeo de campos entre la fuente y el destino (tabla de dimensión *dim_perfil*). Para esto realizará una pequeña transformación en el campo *id_perfil* donde se registrará una secuencia numérica, y por último hará el mapeo de los campos *Edad, Genero y Gusto*:

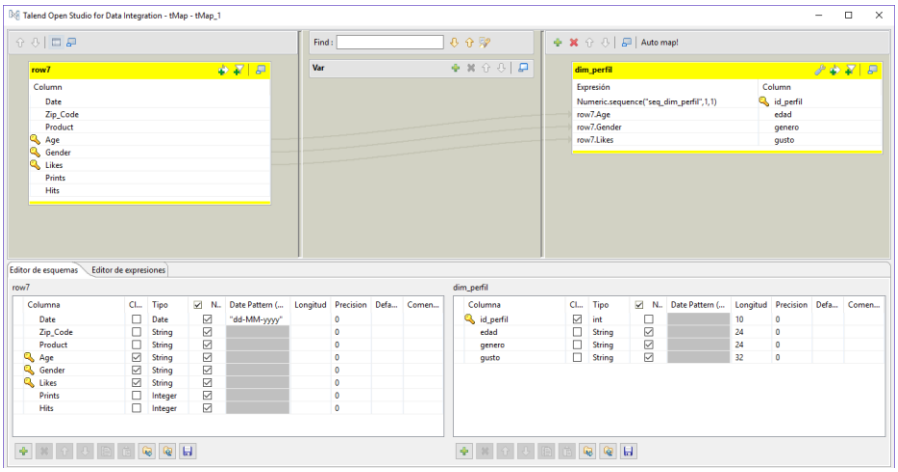


Figura 46. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla *dim_perfil* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

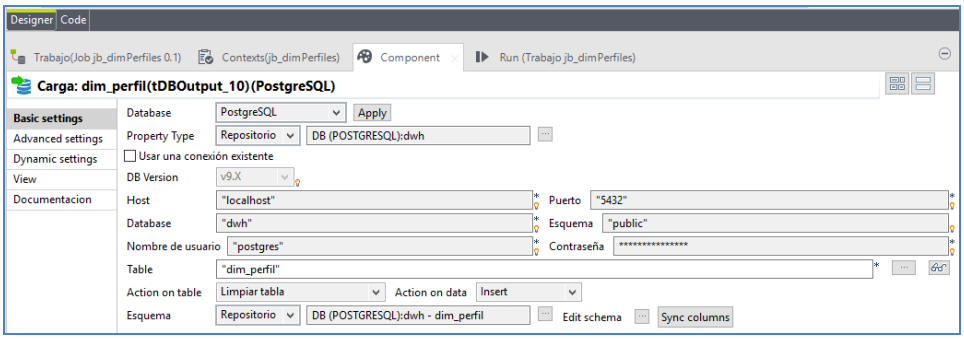


Figura 47. Componente tDBOutput para el cargue de datos en la tabla dim_perfil.

El proceso ETL se ejecuta dando clic en el botón ***Run***:

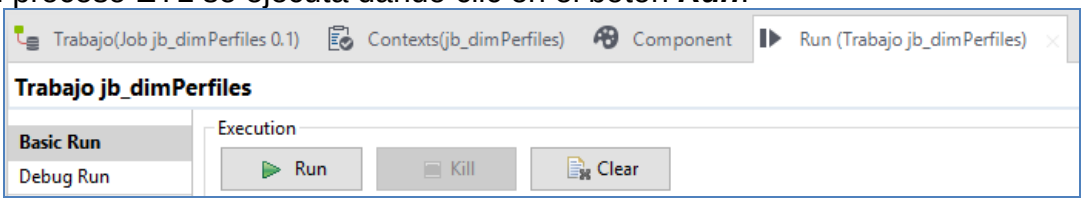


Figura 48. Ejecución proceso ETL jb_dimPerfiles.

4.6. Extracción, transformación y carga de la tabla de hechos

Para cargar los datos en la tabla de hechos que se usará como la tabla neural del cubo a partir del cual se desarrollaran las analíticas para la campaña de marketing digital, se modeló en *Talend Open Studio* un último proceso ETL llamado ***jb_hchAnuncios***. El job extrae los datos a partir de las diferentes fuentes en Excel (*plt_Facebook.xlsx, plt_Instagram, plt_Twitter, plt_Youtube, Productos.xlsx y Zonas.xlsx*) y los carga en la tabla de hechos *hch_anuncio* en el Data Warehouse dentro de la base de datos *PostgreSQL*:

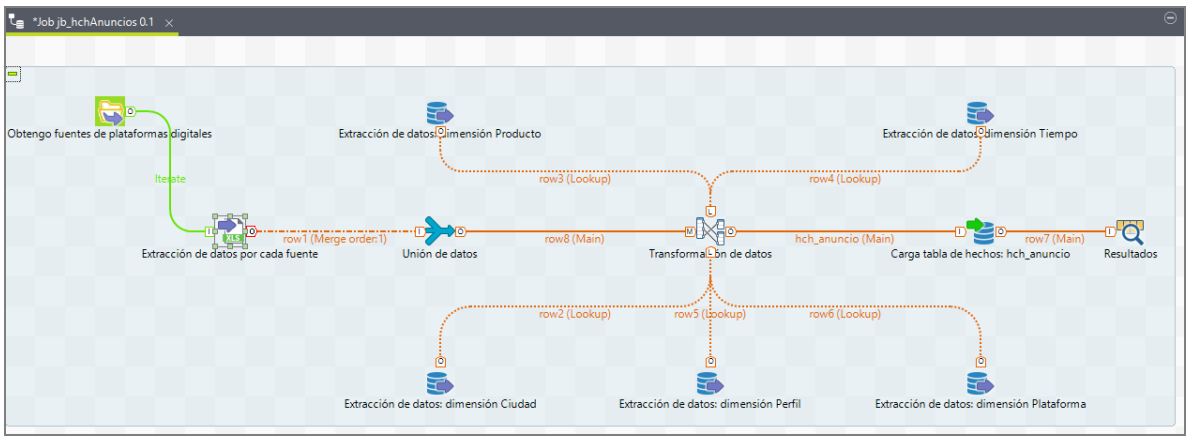


Figura 49. job jb_hchAnuncios implementado en Talend Open Studio.

El proceso inicia con un componente ***tFileList*** que a partir de una carpeta de archivos obtiene el listado de aquellos que cumplan con el patrón de un archivo Excel proporcionado por una plataforma digital: comienzan con el prefijo *"plt_"* y terminan en una extensión *".xlsx"*:

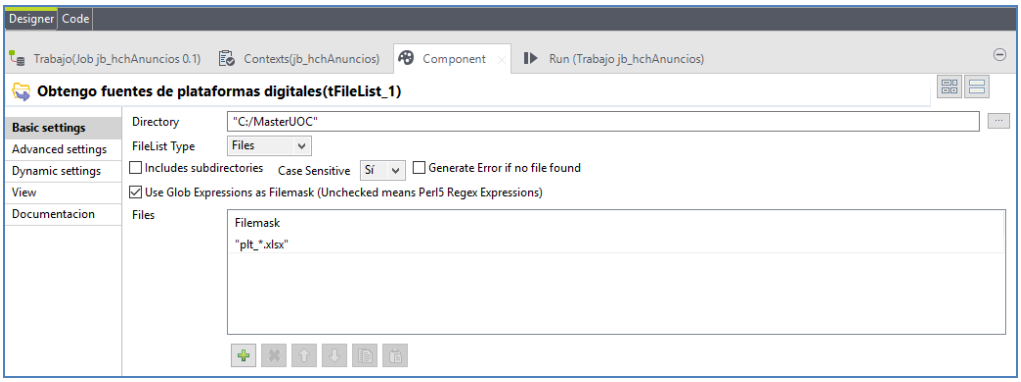


Figura 50. Componente tFileList para listar los archivos plt_[NOMBRE-PLATAFORMA].xlsx

En cada iteración, se obtendrá un archivo de una plataforma digital, el cual será abierto a través del componente ***tFileInputExcel*** y de su contenido se obtendrán todos los campos relacionados con el comportamiento a la campaña publicitaria a través de plataformas digitales: *Date*, *Zip_Code*, *Product*, *Age*, *Gender*, *Likes*, *Plataforma*, *Prints* y *Hits*.

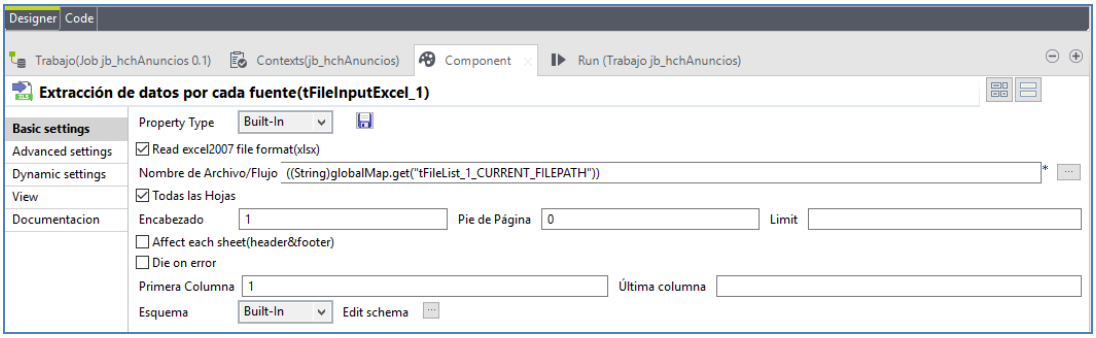


Figura 51. Extracción de los datos de las fuentes proporcionadas por las plataformas digitales.

La salida de cada iteración de archivo de plataforma digital se unirá a través del componente ***tUnit*** cuya salida se conectará como entrada del componente ***tMap***.

Adicional, a través del uso de componente ***tDBInput*** se establecerán cuatro (5) conexiones a las tablas de dimensiones (*dim_tiempo, dim_perfil, dim_plataforma, dim_producto y dim_ciudad*) para a través de "joins" cruzar los valores dimensionales con sus identificadores en las tablas de dimensión y así armar la tabla de hechos, que es el elemento neural del cubo:

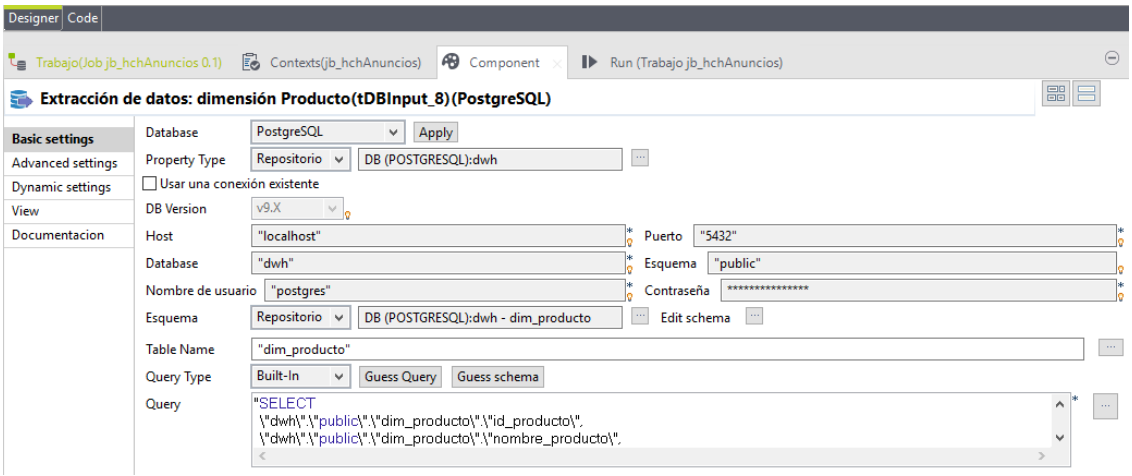


Figura 52. Componente tDBInput para establecer conexiones a tablas de PostgreSQL.

Todos estos flujos de datos (variable, contenido de Excel y contenido de tablas de dimensiones) son enviados como entrada al componente ***tMap*** que se encargará de realizar el mapeo de los campos con los códigos de identificación en las tablas de dimensión fuente y llevarlos a los campos destino en la tabla de hechos *hch_anuncio*, esto a través de relaciones tipo join entre las fuentes, como se muestra en la siguiente imagen. Además de realizar una pequeña transformación en el campo *id_anuncio* donde se registrará una secuencia numérica.

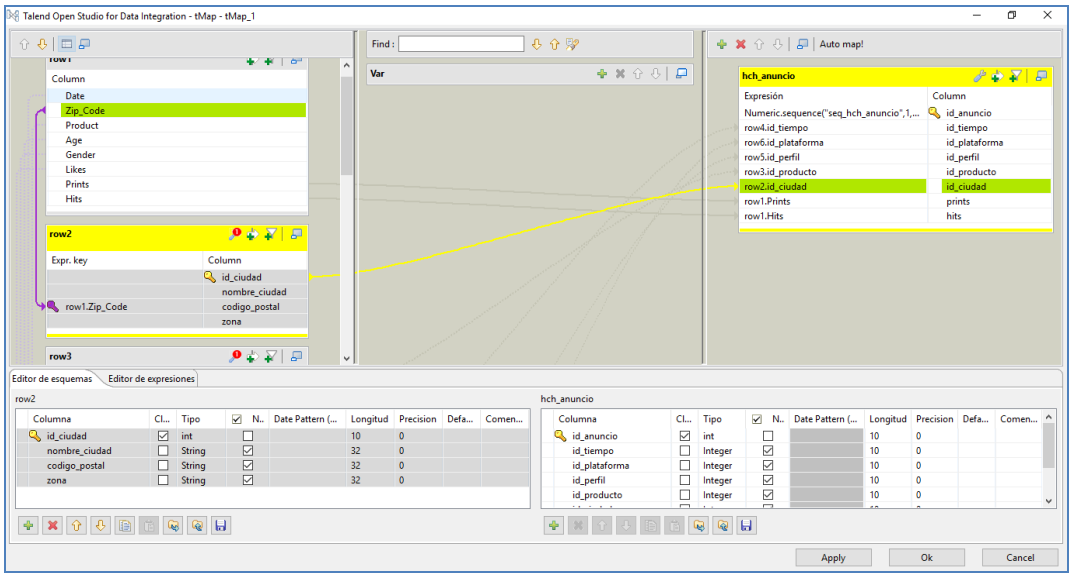


Figura 53. Componente tMap para el mapeo y transformación de datos previo al cargue.

Por último, conectamos la salida de los datos procesados y transformados a un componente ***tDBOutput*** para cargar los datos. En este paso primero se limpia la tabla de hechos *hch_anuncio* antes de realizar la inserción o cargue de los datos procesados en el paso anterior:

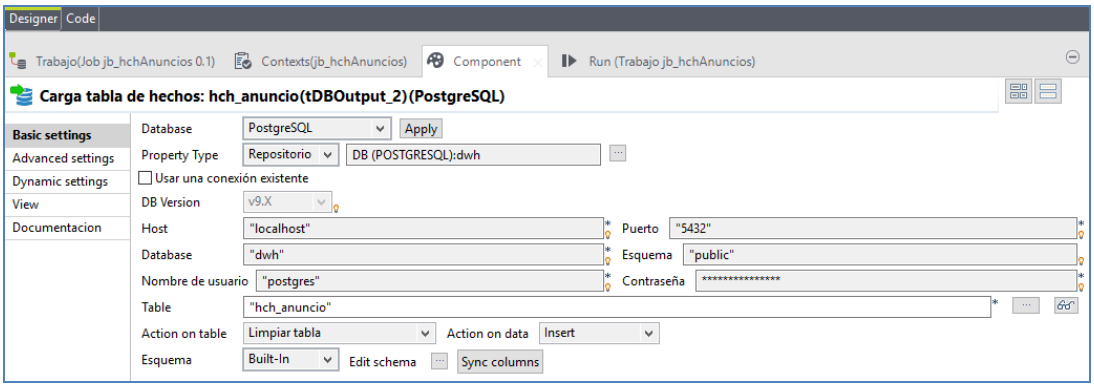


Figura 544. Componente tDBOutput para el cargue de datos en la tabla hch_anuncio**.**

El proceso ETL, para poblar la tabla de hechos, se ejecuta dando clic en el botón ***Run***:

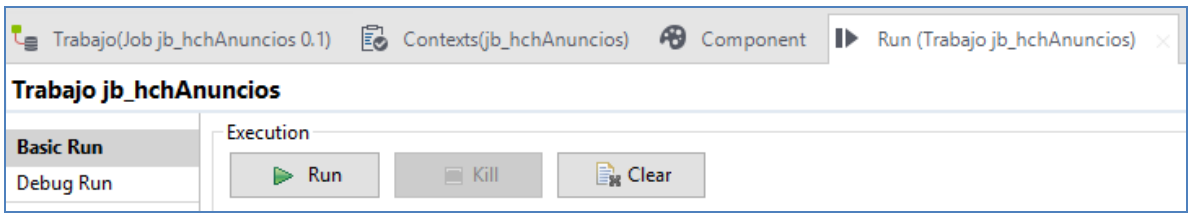


Figura 55. Ejecución proceso ETL jb_hchAnuncios.

5. Implementación del Entorno BI para el análisis

Una vez que se logra la extracción, transformación y carga de los datos dentro del Data Warehouse, se tiene la base de información sobre la cual se pueden diseñar analíticas y visualizaciones que nos permitan explicar algunos comportamientos, responder algunas preguntas y explicar nuestras conclusiones.

La herramienta que se usará como entorno BI para la analítica y visualización de los datos es Microsoft Power BI Desktop, que proporciona un conjunto de herramientas de análisis empresarial que permiten conectarnos a cientos de orígenes de datos, preparación de datos simplificada, generación de análisis ad hoc y estilizados informes interactivos.

5.1. Montaje del entorno BI

Pensando en un entorno corporativo, la solución de Microsoft para inteligencia de negocios es ideal y bastante adaptable.

Power BI ofrece funcionalidades aptas para empresas, y se encuentra potenciada para un modelado exhaustivo y un análisis en tiempo real, con grandes posibilidades para escalarla dentro de las empresas partiendo de una herramienta personal para la creación de informes y visualización, hasta convertirse en el motor de análisis y de decisión que impulsa proyectos estratégicos de grupos de trabajo, unidades de negocio o la corporación entera.

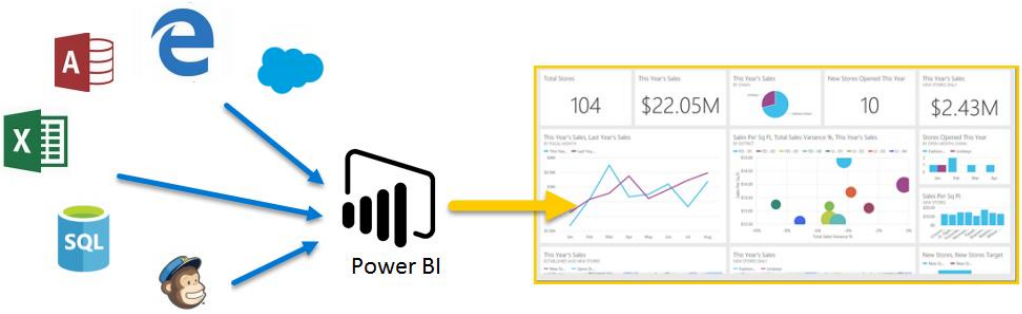


Figura 56. Características Power BI

Power BI es una colección de servicios de software, que consta de una aplicación de escritorio de Windows denominada ***Power BI Desktop***, un servicio por suscripción tipo SaaS (software como servicio) en línea denominado ***Servicio Power BI***, un servidor web On-Premise conocido como Power BI Report Server y aplicaciones móviles de Power BI disponibles para teléfonos y tabletas Windows, así como para dispositivos iOS y Android. Para la práctica enfocada en este trabajo de grado, sólo emplearemos *Power BI Desktop* que será suficiente para implementar las analíticas y visualizaciones requeridas.

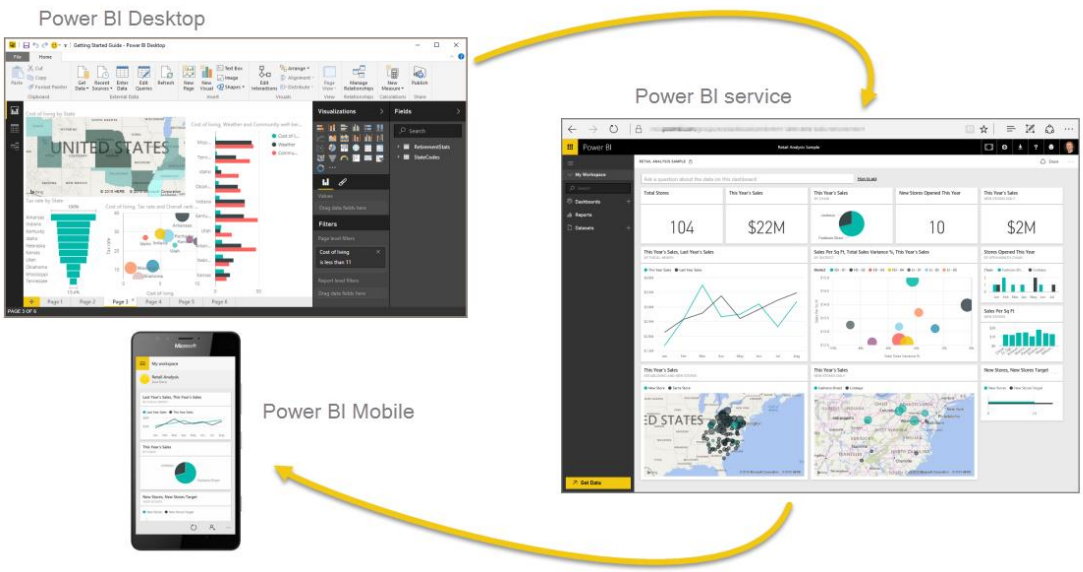


Figura 57. Modalidades Power BI

La aplicación de Power BI en los entornos corporativos varía dependiendo del caso de uso y la madurez de la empresa en procesos y su estructura organizacional alrededor de la inteligencia de negocios.

Por ejemplo, podría darse el caso de que la mayoría de las personas dentro de la empresa consuman los datos y las analítcas a través de Servicio Power BI, mientras que el equipo técnico y de planeación, dedicado a procesar las cifras y crear informes empresariales, use habitualmente *Power BI Desktop* y el resultado de su trabajo sea publicable como informes de Desktop en el *Servicio Power BI* o el *Power BI Report Server*, para la colaboración y el consumo masivo de los datos e indicadores. Mientras el president, la fuerza de ventas o los gerentes, podrían preferir el uso del app móbile de Power BI para supervisar el progreso de sus cuotas de venta y profundizar en los detalles de los nuevos clientes potenciales. Power BI es una solución flexible, adaptable y versátil.

En la siguiente lista, se describen las especificaciones técnicas mínimas para desplegar Power BI Desktop en un ambiente local:

- Windows 7 y Windows Server 2008 R2 o posterior
- .NET 4.5
- Internet Explorer 9 o posterior o navegador compatible
- Memoria (RAM): Al menos 1 GB disponible; se recomienda 1,5 GB o más.
- Pantalla: se recomienda al menos 1440 x 900 o 1600 x 900 (16:9). No se recomiendan las resoluciones inferiores a 1024 x 768 o 1280 x 800, ya que ciertos controles (por ejemplo, para cerrar la pantalla de inicio) solo se muestran en resoluciones superiores a esta.
- CPU: 1 GHz o superior; se recomienda un procesador de x86 o x64 bits.

5.2. Instalación de Power BI

Descargamos la solución gratuita de Power BI Desktop. La versión empresarial te permite una funcionalidad extendida, no incorporada en la versión gratuita, que facilita la publicación de dashboards interactivos en un portal web colaborativo.

Antes de usarlo, y debido a que nuestro Datawarehouse se encuentra en una base de datos PostgreSQL, debemos instalar el driver de conexión llamado NpgSQL v3.2.7. Tener en cuenta incluir dentro de la instalación el componente Npgsql GAC, como se muestra en la siguiente imagen:

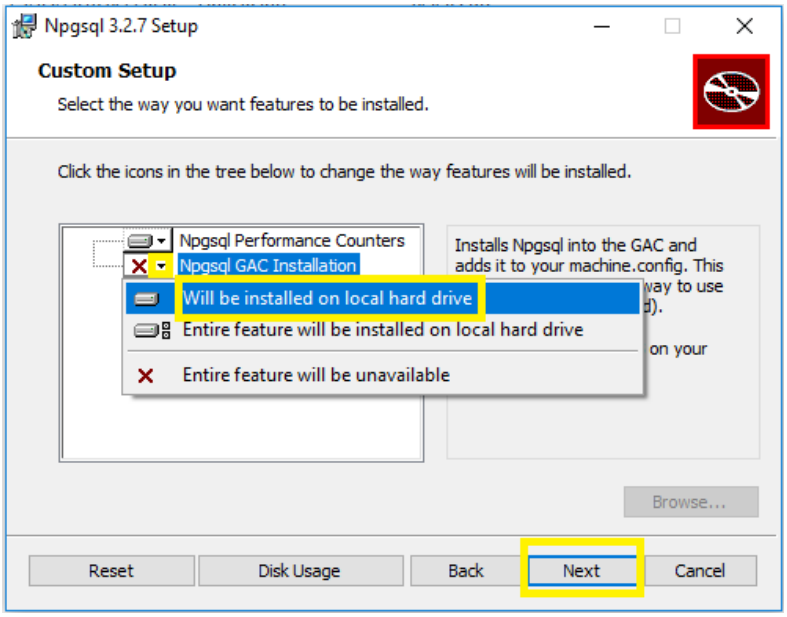


Figura 58. Instalación del driver de conexión para base de datos PostgreSQL.

Antes de abrir Power BI y posterior a la instalación del driver de conexión de PostgreSQL, se recomienda realizar un reinicio del equipo para que se complete la instalación de los demás complementos.

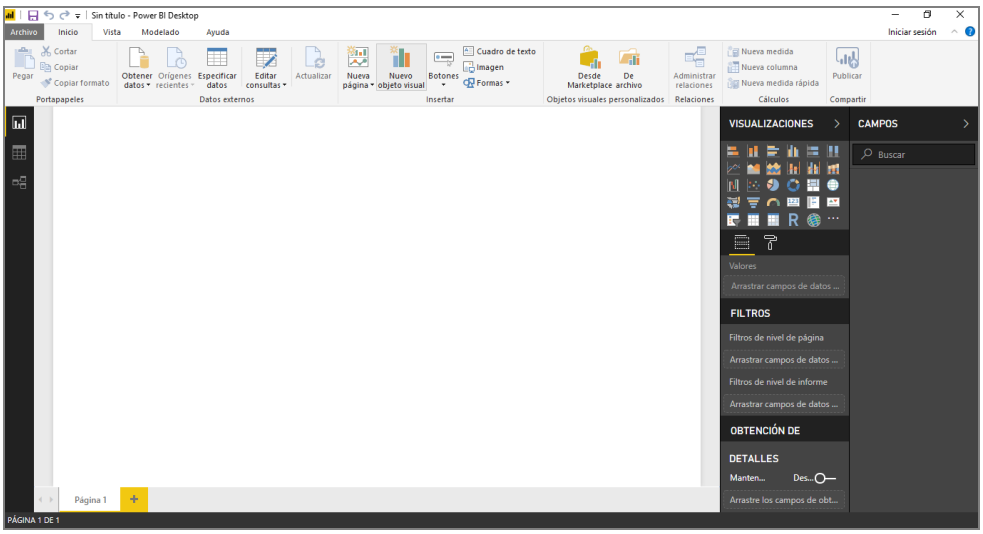


Figura 59. Interfaz de Power BI Desktop.

5.3. Creación del cubo de datos en PowerBI

Lo primero que haremos es establecer una conexión al Data Warehouse, que en nuestro caso reposa en una base de datos PostgreSQL, siguiendo los cuatro (4) pasos indicados a continuación:

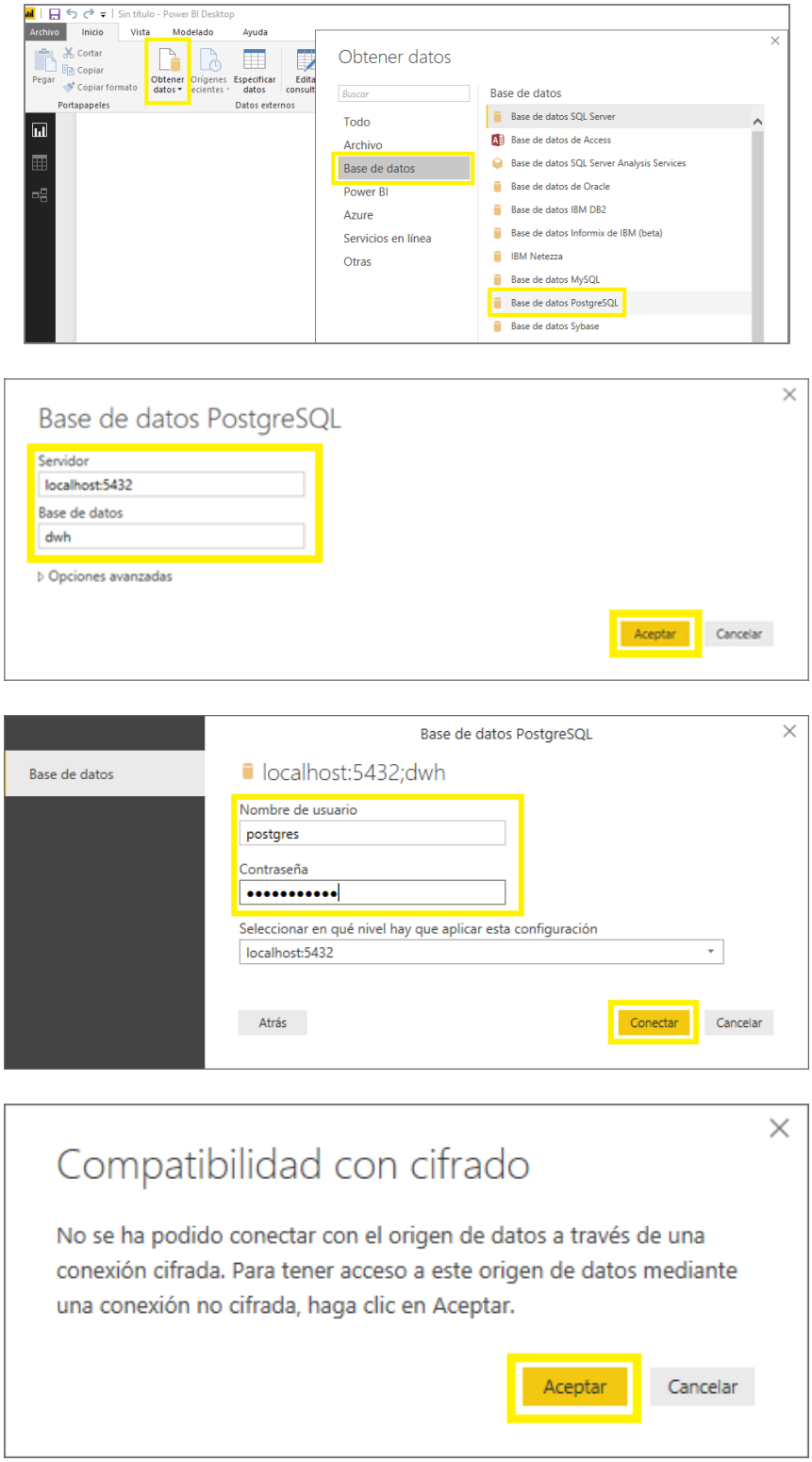


Figura 60. Conexión al Datawarehouse en la base de datos PostgreSQL.

Una vez establecida la conexión, seleccionamos la tabla de hechos y todas las tablas de dimensiones, con las cuales armaremos el cubo de datos.

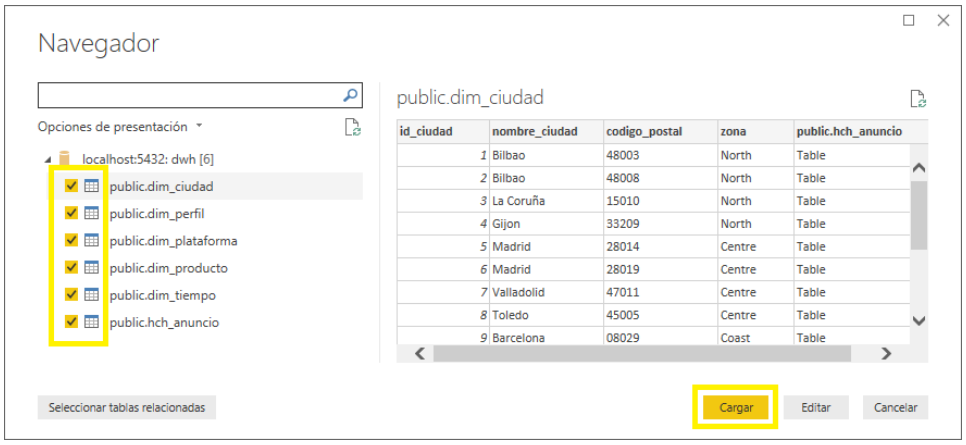


Figura 61. Cargue de tablas de hechos y dimensiones en Power BI.

Una vez cargadas las tablas, podemos dirigirnos a la etiqueta Relaciones y observar el cubo armado.

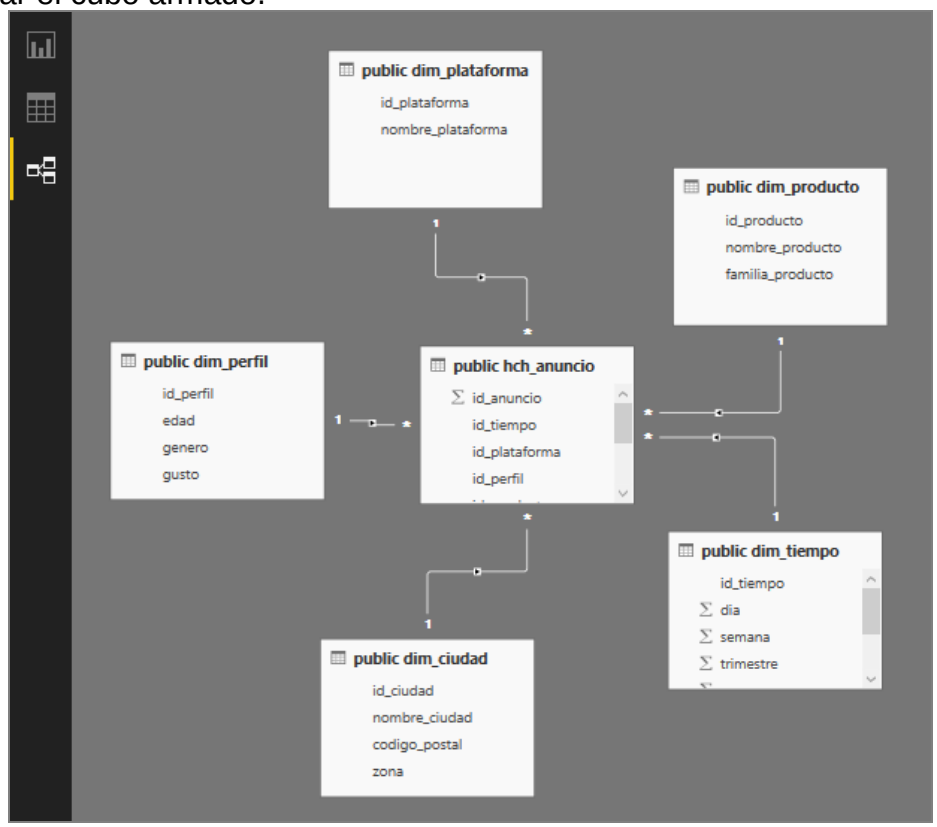


Figura 62. Cubo de datos en Power BI.

Como un dato agregado al cubo, debemos calcular el ***Click Through Rate*** (*CTR*) que es un indicador clave, calculado como un porcentaje, y el cual mide la tasa de efectividad de anuncios o publicidad a través de medios digitales, el cual se obtiene de la siguiente fórmula:

CTR = (Impresiones / Clics) x 100%

En Power BI, este dato lo obtendremos creando un campo calculado, como se muestra en la siguiente imagen:

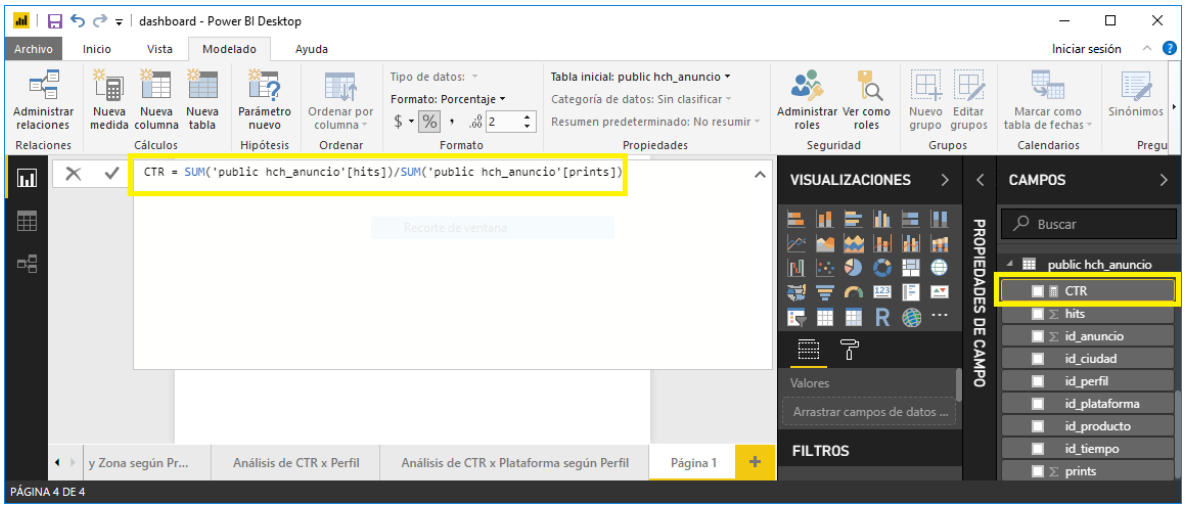


Figura 63. Campo calculado en Power BI para obtener el indicador CTR.

6. Análisis de la información

Una vez que se construye el cubo en Power BI, el siguiente paso es el diseño de las visualizaciones, las analíticas de datos y la unión de todos estos elementos en un cuadro de mando que permita sacar algunas conclusiones.

6.1. Preguntas analíticas

A continuación, listaremos las preguntas sobre las cuales basaremos la analítica y la construcción de las visualizaciones en Power BI:

- ¿Qué regiones o ciudades tienen mejores indicadores de efectividad?
- ¿Hay alguna relación con el producto o familia de productos?
- ¿Existen alguna relación entre la mejora de los indicadores de efectividad con algún segmento de la población objetivo?
- ¿Hay alguna plataforma donde, bajo las mismas condiciones, se obtengan mejores tasas de visualización?
- ¿Existen relaciones entre plataformas y franjas de edad de usuarios que provoquen mejores tasas de visualización?
- ¿El conocimiento del grupo de interés de los usuarios podría ayudar a mejorar los indicadores para determinados productos?

6.2. Análisis por región y ciudad

Buscando responder las preguntas: *¿Qué regiones o ciudades tienen mejores indicadores de efectividad?* y si *¿hay alguna relación entre la zona geográfica con el producto o familia de productos?*, creamos un dashboard en Power BI que incluya una visualización en mapa con un mapa de color, donde el color más intenso corresponda a una tasa de efectividad (CTR) mayor y el color más tenue a un indicador de efectividad mucho menor. Adicional, incluimos una tabla jerárquica de zonas que puede desplegarse para llegar al nivel de ciudades, ordenada de forma descendente por el valor del indicador CTR y otra tabla que, de acuerdo con la elección de la zona o ciudad, o un conjunto de ciudades del mapa, nos indique el ranking de productos con mejor efectividad en anuncios sobre plataformas digitales.

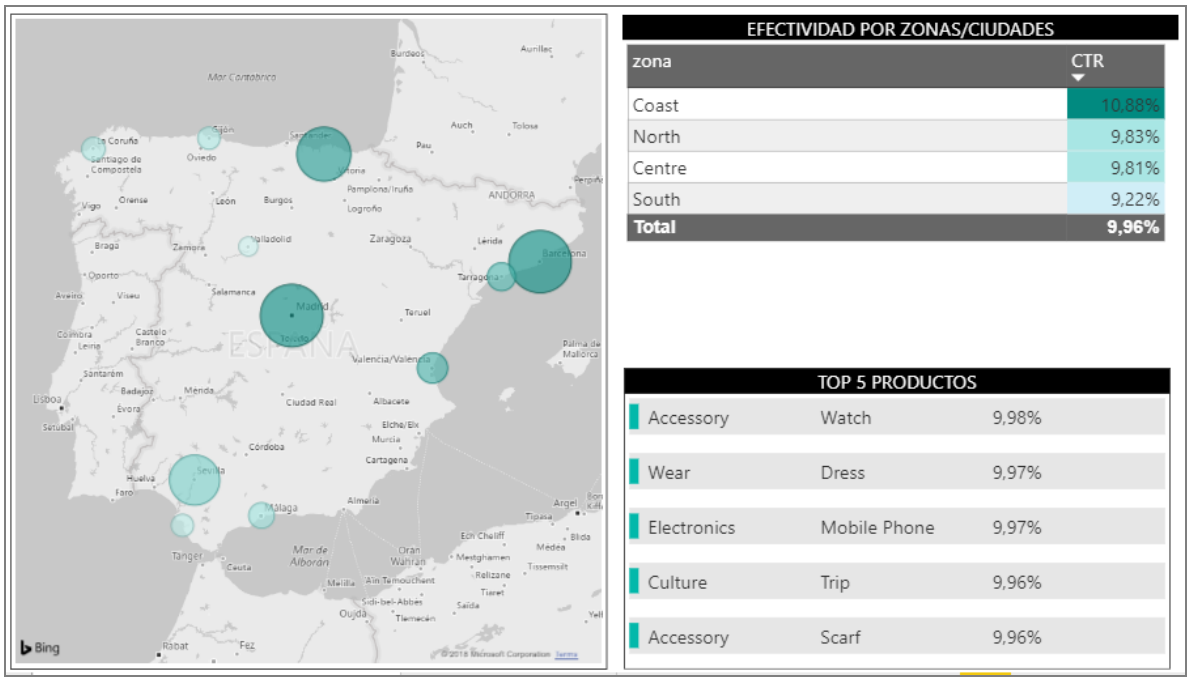


Figura 64. Dashboard con análisis de efectividad por ubicación y productos.

Si observamos la tabla de zonas, podemos concluir que, aunque la zona Costera de España parece tener mejor respuesta a la publicidad a través de canales digitales, parece que no hay una ciudad dentro de esta zona que sea ampliamente dominante, pues observando el mapa vemos que Barcelona se equipará a otros grandes centros urbanos como Madrid (Centro) y Bilbao (Norte) como las ubicaciones con mayor tasa de efectividad CTR.

Otro aspecto para observar es el tamaño de las esferas en la visualización del mapa, atributo que hemos ligado al número total de *"prints"* realizados, es decir, al número total de apariciones de los anuncios en las plataformas mientras los usuarios navegan. Podemos concluir que una mayor exposición de los anuncios no necesariamente determina una mayor conversión en clics, pues Tarragona y Valencia se acercan a unos indicadores *CTR* altos sin mayor exposición de anuncios, contrario a Sevilla cuya respuesta de clics frente a un número alto de impresiones es menos efectiva.

Si seleccionamos los tres (3) centros urbanos con mejor indicador CTR y analizamos el ranking de efectividad por productos y familias de productos, podemos concluir que la ropa y los accesorios para ropa son los que jalonan estos indicadores, muy seguidos por las actividades de corte cultural.

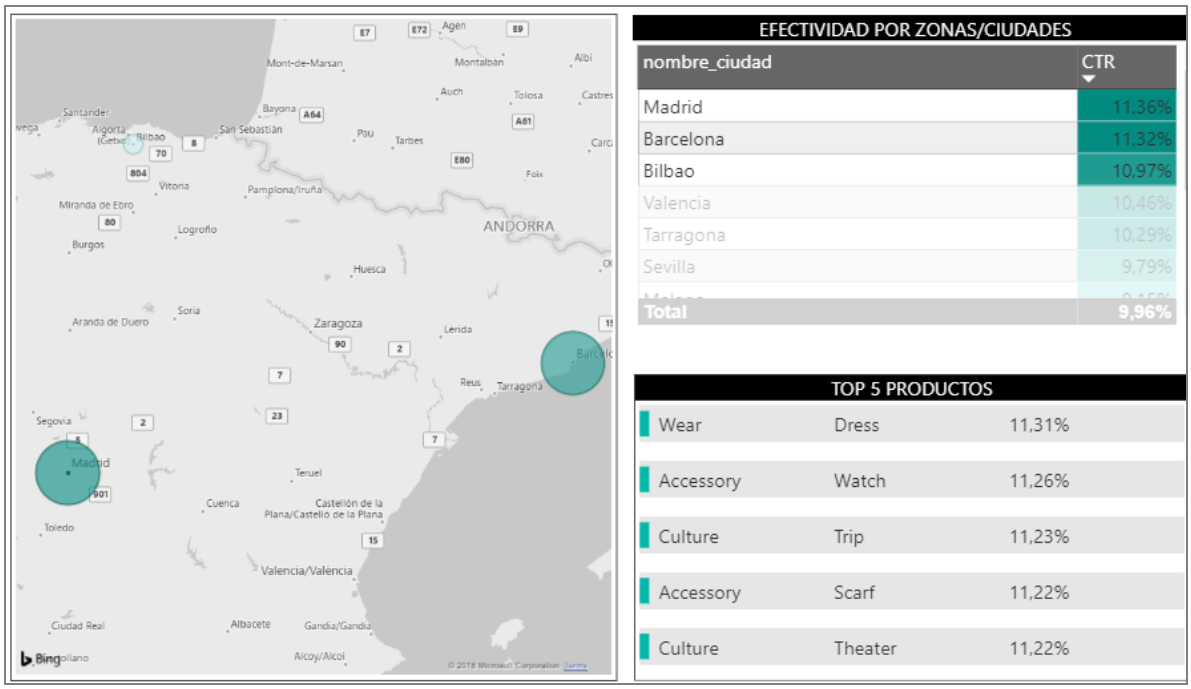
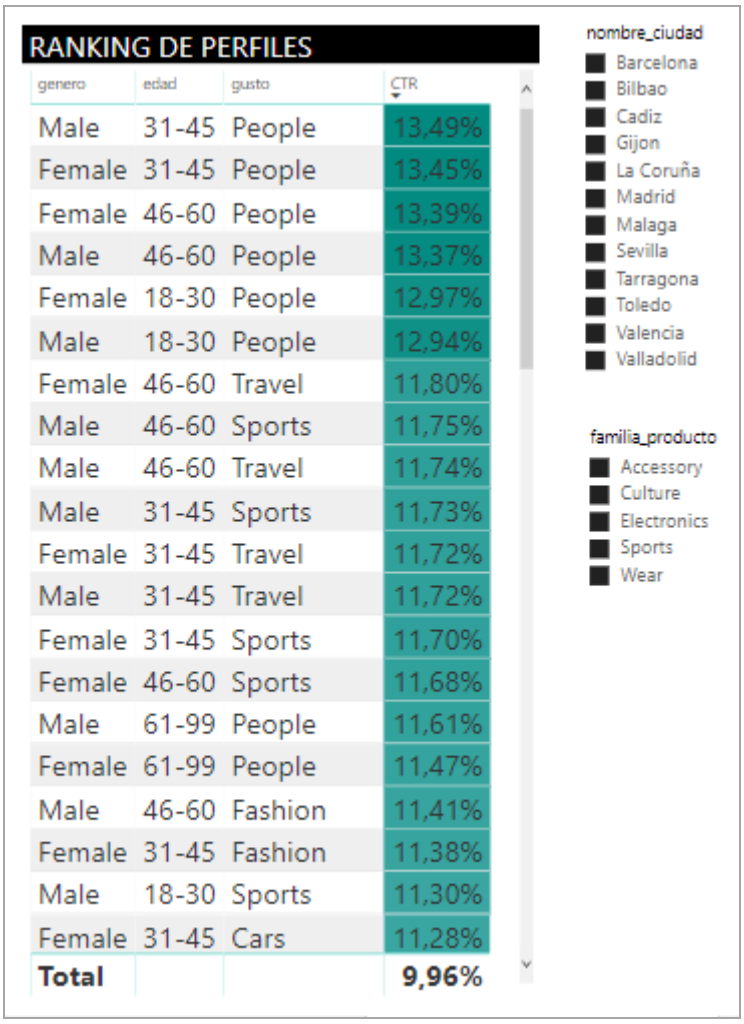


Figura 65. Análisis de productos más efectivos en las ciudades con mejor CTR.

6.3. Análisis por segmento de la población objetivo

Ahora lo que sigue es determinar cuáles son las características de las personas que responden positivamente a estos anuncios, *¿existe alguna relación entre la mejora de los indicadores de efectividad con algún segmento de la población objetivo?*

Si observamos el Ranking de Perfiles relacionados con una respuesta positiva a la publicidad digital, sin distinción de ubicación geográfica ni tipo de producto involucrado en su interacción, no encontramos una mayor distinción en el género de las personas, con una participación equitativa entre hombres y mujeres; sin embargo en rango de edad parece haber una tendencia en las personas que oscilan entre los 31 y los 60 años, donde predomina de forma altamente marcada un gusto por las temáticas sociales (Likes = People).



RANKING DE PERFILES

genero	edad	gusto	CTR
Male	31-45	People	13,49%
Female	31-45	People	13,45%
Female	46-60	People	13,39%
Male	46-60	People	13,37%
Female	18-30	People	12,97%
Male	18-30	People	12,94%
Female	46-60	Travel	11,80%
Male	46-60	Sports	11,75%
Male	46-60	Travel	11,74%
Male	31-45	Sports	11,73%
Female	31-45	Travel	11,72%
Male	31-45	Travel	11,72%
Female	31-45	Sports	11,70%
Female	46-60	Sports	11,68%
Male	61-99	People	11,61%
Female	61-99	People	11,47%
Male	46-60	Fashion	11,41%
Female	31-45	Fashion	11,38%
Male	18-30	Sports	11,30%
Female	31-45	Cars	11,28%
Total			**9,96%**

Figura 66. Análisis del perfil objetivo.

Si analizamos el comportamiento de las personas que responden positivamente a los anuncios teniendo en cuenta tan solo los tipos de productos que persiguen en sus interacciones a través de las diferentes plataformas digitales, encontramos una correlación directa con sus gustos. También se puede concluir que este comportamiento es constante en todas las regiones de España:

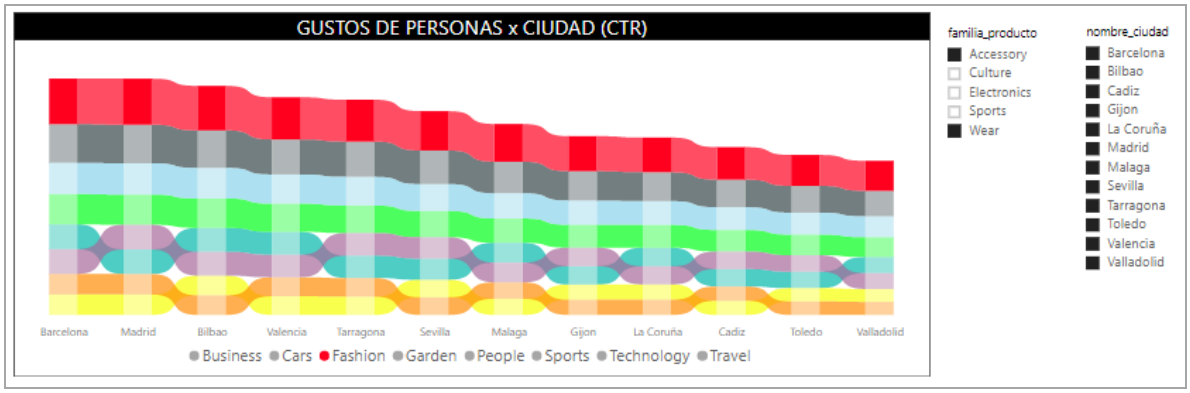


Figura 67. Análisis de gustos de personas que buscan ropa y accesorios.

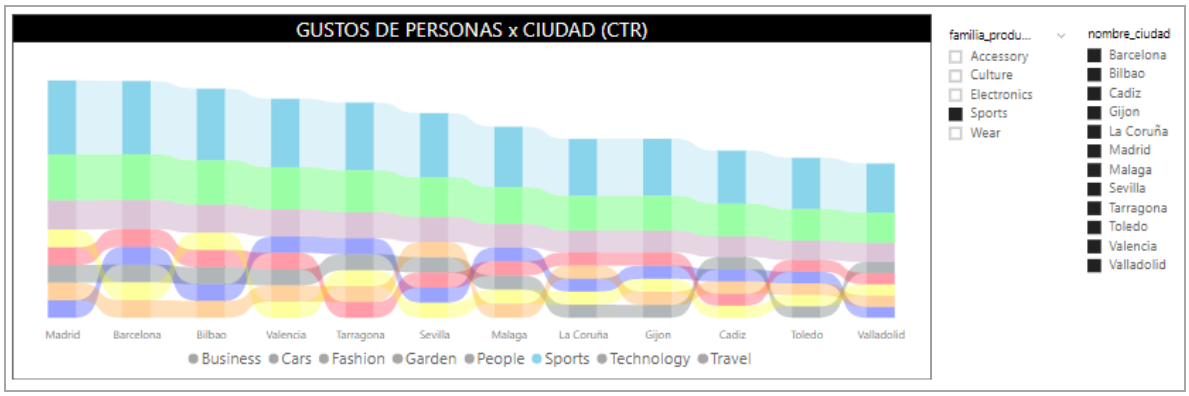


Figura 68. Análisis de gustos de personas que buscan artículos deportivos.

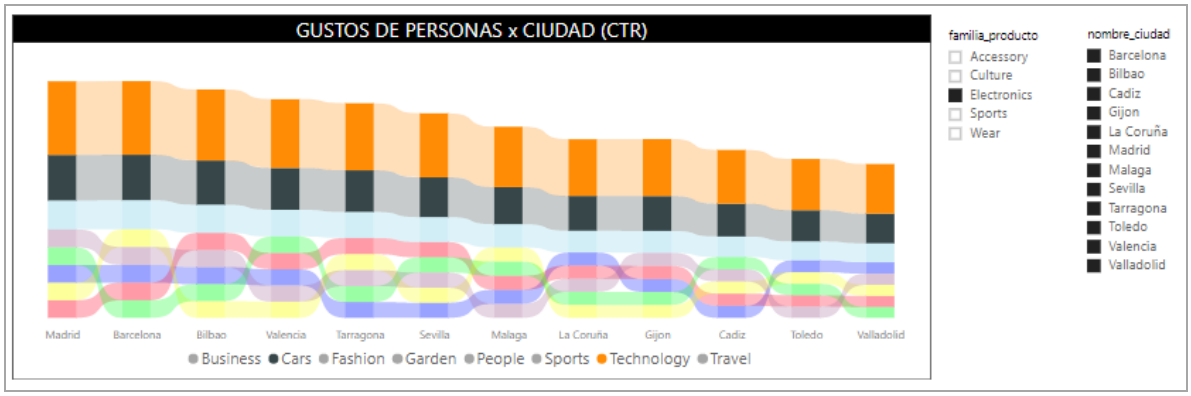


Figura 69. Análisis de gustos de personas que buscan artículos electrónicos.

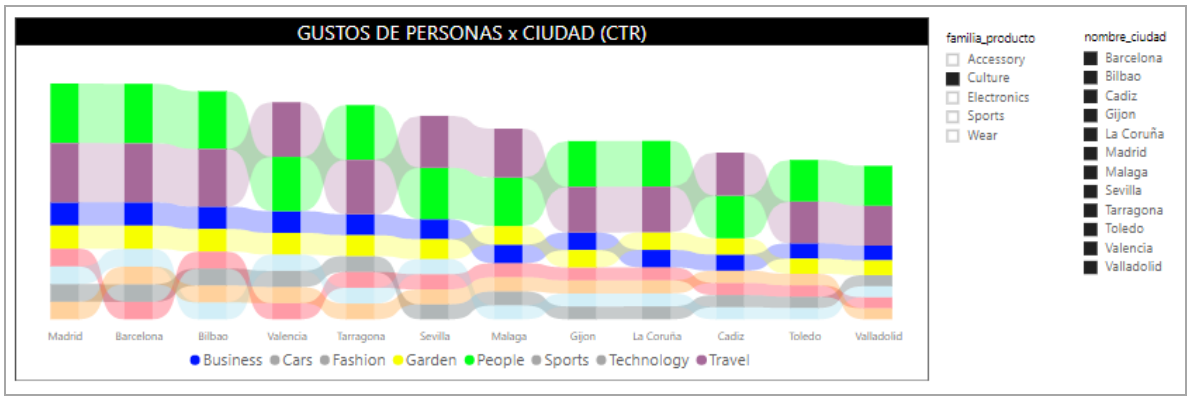


Figura 70. Análisis de gustos de personas que buscan productos/servicios de corte cultural.

Cuando realizamos el análisis por rango de edades, encontramos que la mejor respuesta a los anuncios en las dos grandes urbes parece estar concentrada en adultos (31-45); mientras que, si nos enfocamos en los consumidores potenciales de artículos deportivos y ropa, el grupo de interés cambia a los adultos mayores (46-60), salvo en Valencia donde se conserva los adultos como población objetivo:

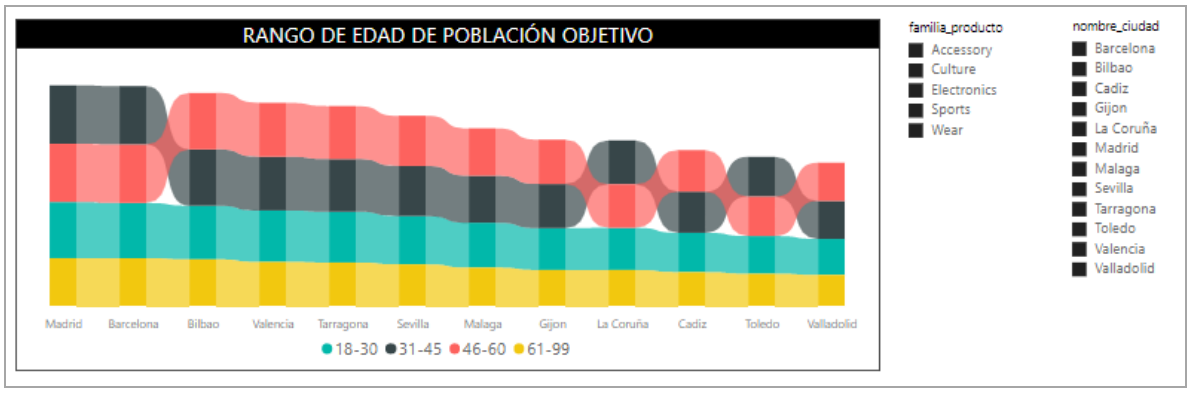


Figura 71. Análisis de rangos de edad de personas de población objetivo.

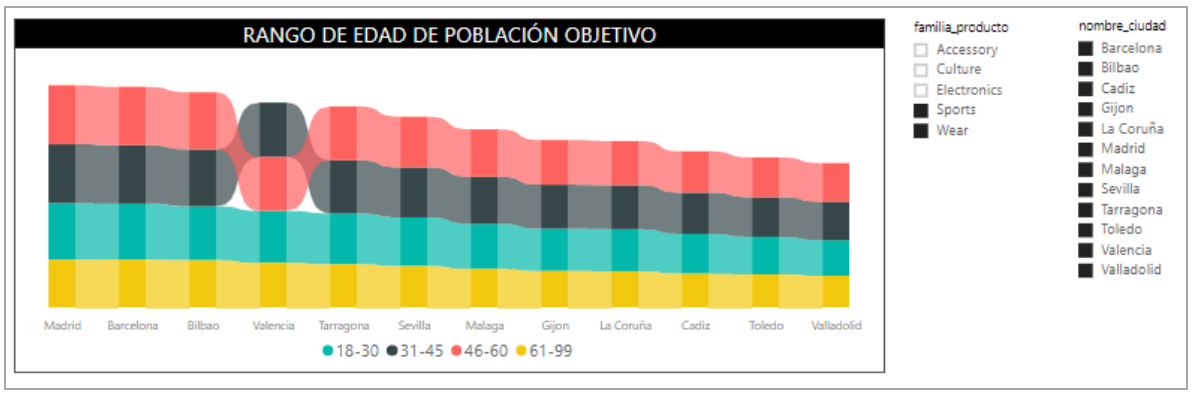


Figura 72. Análisis de rangos de edad de personas que buscan productos deportivos y ropa.

En este orden de ideas, ¿el conocimiento del grupo de interés de los usuarios podría ayudar a mejorar los indicadores para determinados productos? La respuesta es: por supuesto que sí. El género no resulta significativo, pero el gusto de la persona es una variable altamente conexa al tipo de productos publicitados y la edad puede subordinar el consumo en artículos específicos como la moda y el calzado.

6.4. Análisis por plataforma

Si realizamos el análisis por tipo de plataforma digital y lo validamos mediante una línea de tiempo, encontramos un comportamiento bastante homogéneo en la tasa de conversión de publicidad.

Lo interesante es observar los datos en las dos grandes urbes (*Madrid* y *Barcelona*) cuando los anuncios corresponden a publicidad de artículos deportivos y de electrónica desplegados tan sólo en *Instagram* y *YouTube*, donde hay una alta tasa de conversión en clics muy coherente con los gustos y para un segmento exclusivo de adultos jóvenes (18-30).

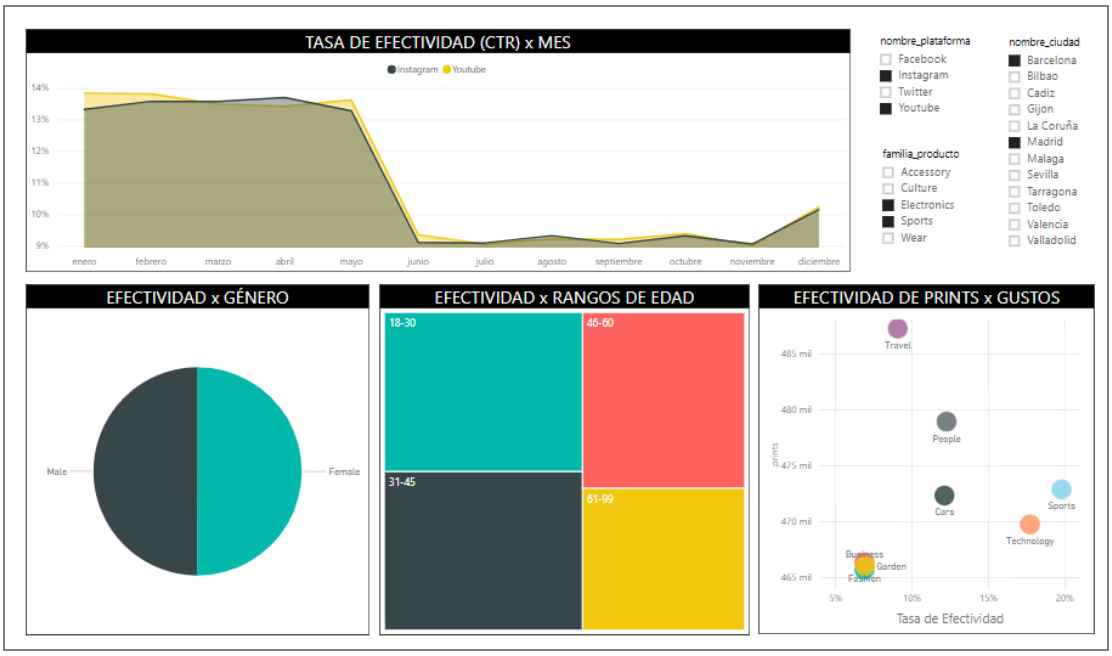


Figura 73. Análisis de efectividad de la publicidad en el segmento de adultos jóvenes.

Si repetimos el ejercicio, pero en esta ocasión escogiendo un tipo de producto de tipo cultural, observamos que la tasa de conversión de la publicidad es altísima en plataformas digitales de opinión como YouTube, Facebook y Twitter, en personas maduras que oscilan entre los 46 y 60 años, cuyos gustos están bastante marcados hacia lo social (People) y el turismo (Travel).

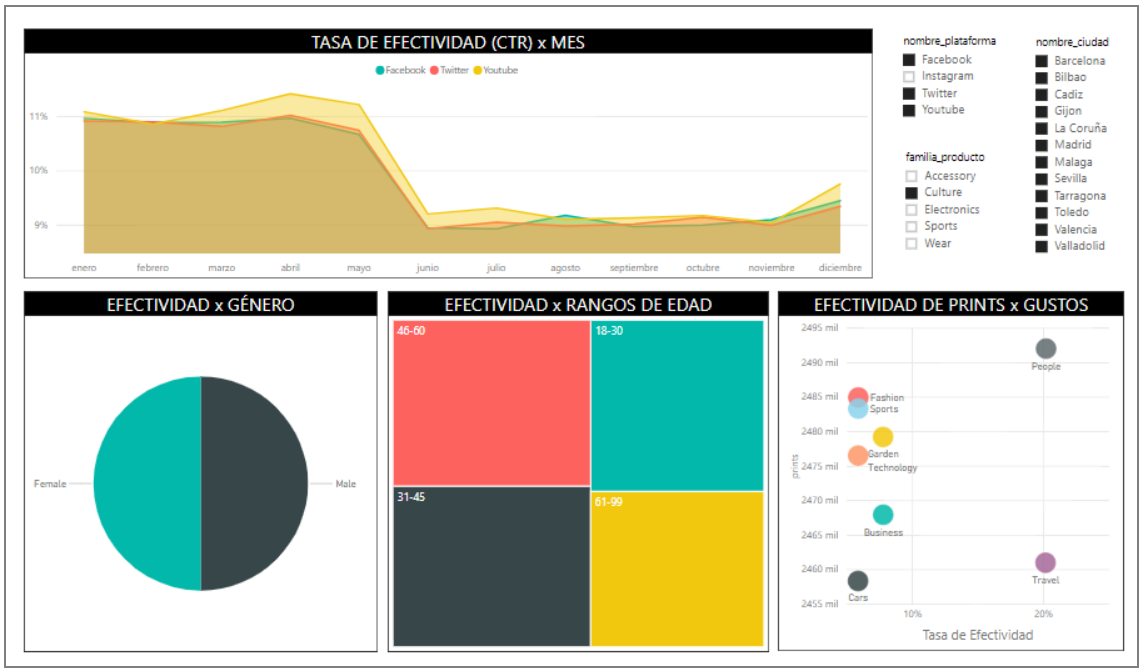


Figura 74. Análisis de efectividad de la publicidad en el segmento de personas maduras.

Ahora, sí analizamos el segmento de productos que mejor efectividad en publicidad reporta, me refiero a la ropa y los accesorios de moda, dentro de las tres grandes urbes (Madrid, Barcelona y Bilbao), podemos observar que se conserva la hipótesis inicial que indica un rango de edad entre los 31 y 60 años

con gustos por lo fashion, sin distinción de plataformas digitales; sin embargo, si filtramos sólo por las plataformas de Instagram y YouTube, observamos como los adultos jóvenes, con similares gustos por lo fashion, toman el rol protagónico en la respuesta publicitaria y aumentan la tasa de conversión.

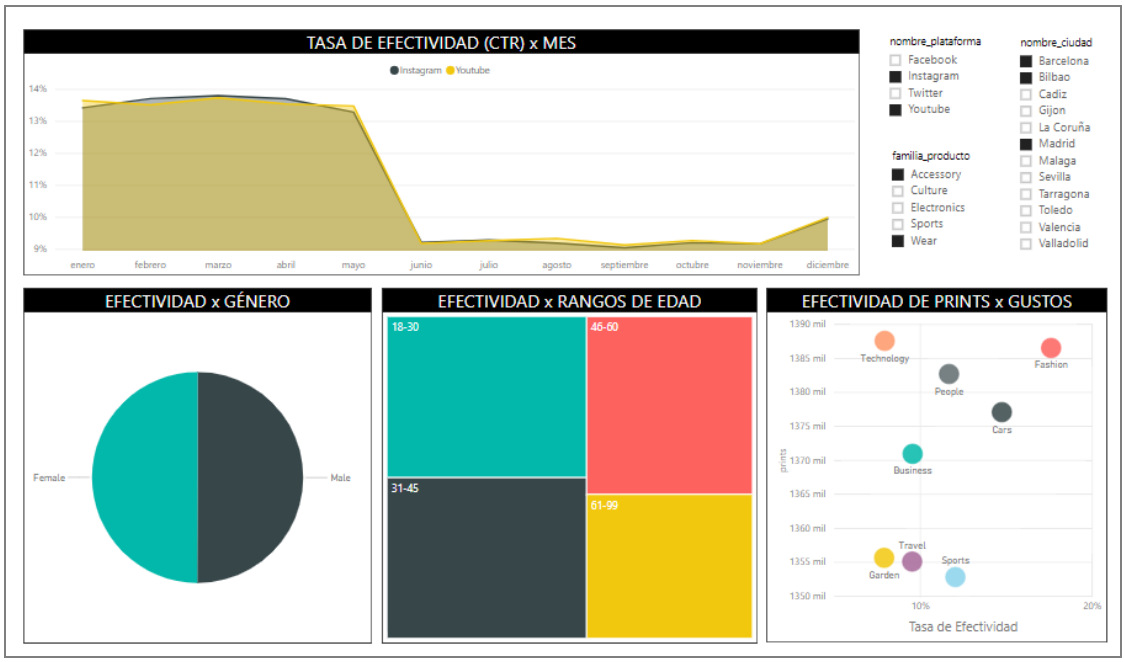


Figura 75. Efectividad de la publicidad de moda en usuarios de Instagram y Youtube.

6.5. Conclusiones de la campaña de anuncios en plataformas digitales

Una vez realizados los cuadros de mando y la analítica de los datos se pueden elaborar y registrar las conclusiones para determinar el grado de efectividad de la campaña de anuncios desplegada a través de plataformas digitales.

Las conclusiones estarán enmarcadas en dar respuesta al cuestionamiento principal: En general, ¿cuál es la parametrización más efectiva de anuncios en las principales plataformas digitales?

Las conclusiones que podríamos extraer del análisis realizado son las siguientes:

- La publicidad en moda tiene una alta conversión en ventas sobre todo en las tres (3) más grandes urbes de España: Barcelona, Madrid y Bilbao.
- Si la publicidad se enfoca en desplegarse en plataformas digitales como Instagram y YouTube se llega con mayor efectividad a usuarios jóvenes con edades entre los 18 y 30 años.
- Si realizamos un análisis por zonas, la conversión en clic es alta y homogénea en las poblaciones de la costa de España.
- En el segmento de personas maduras (46-60), con gustos por lo social y el turismo, la publicidad más efectiva de visualización corresponde a temáticas culturales.

- En el segmento de adultos jóvenes (18-30), con gustos por los deportes y la tecnología, la publicidad más efectiva de visualización corresponde a temáticas afines a sus gustos.
- El género no es un atributo relevante, a diferencia del gusto (Likes) que es determinante en la conversión de la publicidad en clics.
- El perfil que mejor respuesta da a la publicidad a través de plataformas digitales corresponde a hombres entre los 31 y 45 años con gustos por lo social (People).

7. Conclusiones

- El presente trabajo del Master de Inteligencias de Negocio ha sido muy importante y útil porque me ha permitido adquirir conocimientos y experiencia en el área de Inteligencia de Negocios, el cual seleccioné no sólo por la demanda creciente en este tema sino también porque en la empresa donde trabajo van a iniciar desarrollos en esta área.

- Las lecciones aprendidas correspondieron en primera instancia a la arquitectura de BI, la definición y diseño del DWH , así como las ETL, en esta parte la investigación fue amplia porque era importante conocer los diferentes componentes que Talend dispone para la creación de las ETL.

  Por otro lado, la construcción de los cuadros de mando en Power BI también fue otro aprendizaje, sin embargo por tratarse de una herramienta de fácil uso, no sólo permitió un aprendizaje rápido, sino que fue posible realizar un análisis con gran detalle sobre los datos.

- Existen diversas herramientas Open Source con las que podemos trabajar este tipo de proyectos, sin embargo siempre es importante evaluar las opciones para seleccionar las más adecuadas teniendo en cuenta el hardware, software e incluso los requerimientos de alta gerencia con respecto a la forma de visualización en diversos dispositivos tecnológicos.

- Los objetivos definidos al inicio del proyecto, en el trabajo final de Máster fueron alcanzados y lo más relevante es que se pudo verificar la efectividad de campaña publicitaria basada en el CTR de los anuncios publicados en las distintas plataformas digitales.

- El desarrollo del proyecto fue satisfactorio puesto que se realizó en base a la planificación planteada, no hubo desviación alguna, y en cuanto a la metodología, Scrum permitió producir los diferentes entregables en las fechas indicadas.

- Una línea a explorar en el futuro que no fue contemplado en este proyecto debido al tiempo y a que no se contaba con la licencia requerida, sería la gestión de permisos entre diferentes usuarios, así como la acción de publicar dashboards interactivos en un portal web colaborativo.

8. Glosario

BI: Siglas en Ingles que significa Business Intelligence**.**

Business Intelligence(Inteligencia de Negocios): Conjunto de técnicas y herramientas para transformar los datos e información de una organización en conocimiento con el cual tomar decisiones estratégicas.

Dashboard: Es una representación gráfica de los principales indicadores (KPI) que intervienen en la consecución de los objetivos de negocio, y que está orientada a la toma de decisiones para optimizar la estrategia de la empresa. Un dashboard debe transformar los datos en información y está en conocimiento para el negocio.

Data Mart: Base de datos especializada, que almacena los datos de un área de negocio específica.

Data Warehouse: Almacén de datos (del inglés data warehouse) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

DWH: Siglas en Ingles que significa **Data Warehouse.**

ETL («extraer, transformar y cargar», frecuentemente abreviado **ETL**)**:** es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

9. Bibliografía

[1] https://www.captio.net/blog/la-ventaja-competitiva-un-factor-estrategico-del-bi
[2] https://www.scrumguides.org/docs/scrumguide/v2016/2016-Scrum-Guide-Spanish-European.pdf
[3] https://es.linkedin.com/pulse/kimball-e-inmon-y-el-dise%C3%B1o-de-data-warehouses-william-qui%C3%B1onez
[4] file:///C:/Users/keinthpc/Downloads/TD_BarbaraEgea.pdf
[5] https://es.wikipedia.org/wiki/Cuadro_de_mando_integral
[6] https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/ventajas-en-la-integraci-n-de-datos-con-la-herramienta-etl-pentaho
[7] https://prezi.com/tfzf5ljcnqcq/cloveretl/
[8] https://guiadev.com/5-alternativas-a-mysql-server/
[9] https://guiadev.com/mariadb-vs-mysql-cual-debo-elegir/
[10]https://bddimensionales.wikispaces.com/Modelado+Dimensional
[11]https://bddimensionales.wikispaces.com/Modelado+Dimensional
[12] https://es.wikipedia.org/wiki/Esquema_en_copo_de_nieve

10. Anexos

Id	Nombre de tarea	Duración	Comienzo	Fin
1	**Sistema BI para el análisis de la publicidad en entornos**	76 días	lun 26/02/18	lun 11/06/18
2	**Planificación del Proyecto**	**6 días**	**lun 26/02/18**	**lun 05/03/18**
3	Establecer Contexto y Justificación	1 día	lun 26/02/18	lun 26/02/18
4	Definición de objetivos	1 día	mar 27/02/18	mar 27/02/18
5	Descomposición del trabajo en tareas	1 día	mié 28/02/18	mié 28/02/18
6	Planificación Tareas en Hitos	3 días	jue 01/03/18	lun 05/03/18
7	Entrega PEC1 - Planificación Proyecto	0 días	lun 05/03/18	lun 05/03/18
8	**Análisis y Selección de las herramientas**	**11 días**	**mar 06/03/18**	**mar 20/03/18**
9	Análisis de las herramientas BI	5 días	mar 06/03/18	lun 12/03/18
10	Documentación Memoria	3 días	mar 13/03/18	jue 15/03/18
11	Selección Herramienta para el proyecto	2 días	vie 16/03/18	lun 19/03/18
12	Propuesta Herramienta seleccionada	1 día	mar 20/03/18	mar 20/03/18
13	**Creación del Entorno de Trabajo**	**3 días**	**mié 21/03/18**	**vie 23/03/18**
14	Verificación requisitos de instalación	1 día	mié 21/03/18	mié 21/03/18
15	Instalación Servidor Base de Datos	1 día	jue 22/03/18	jue 22/03/18
16	Instalación del Software BI	1 día	vie 23/03/18	vie 23/03/18
17	**Diseño del Datawarehouse**	**11 días**	**lun 26/03/18**	**lun 09/04/18**
18	Análisis de requerimientos	2 días	lun 26/03/18	mar 27/03/18
19	Modelo Conceptual de Datos	3 días	mié 28/03/18	vie 30/03/18
20	Modelo Lógico de Datos	3 días	lun 02/04/18	mié 04/04/18
21	Modelo Fisico de Datos	3 días	jue 05/04/18	lun 09/04/18
22	Entregable PEC2 - Selección Herramienta BI y Diseño DWH	0 días	lun 09/04/18	lun 09/04/18
23	**Construcción del Datawarehouse**	**4 días**	mar 10/04/18	vie 13/04/18
24	**Diseño ETL**	**10 días**	lun 16/04/18	vie 27/04/18
25	**Construcción ETL**	**5 días**	lun 30/04/18	vie 04/05/18
26	Entregable PEC3 - Diseño y Construcción ETL	0 días	lun 07/05/18	lun 07/05/18
27	**Implementación del Front-End del usuario**	12 días	mar 08/05/18	mié 23/05/18
28	**Análisis Resultados Obtenidos**	8 días	mié 23/05/18	vie 01/06/18
29	**Conclusiones**	5 días	lun 04/06/18	vie 08/06/18
30	Entrega Final del Proyecto	0 días	lun 11/06/18	lun 11/06/18

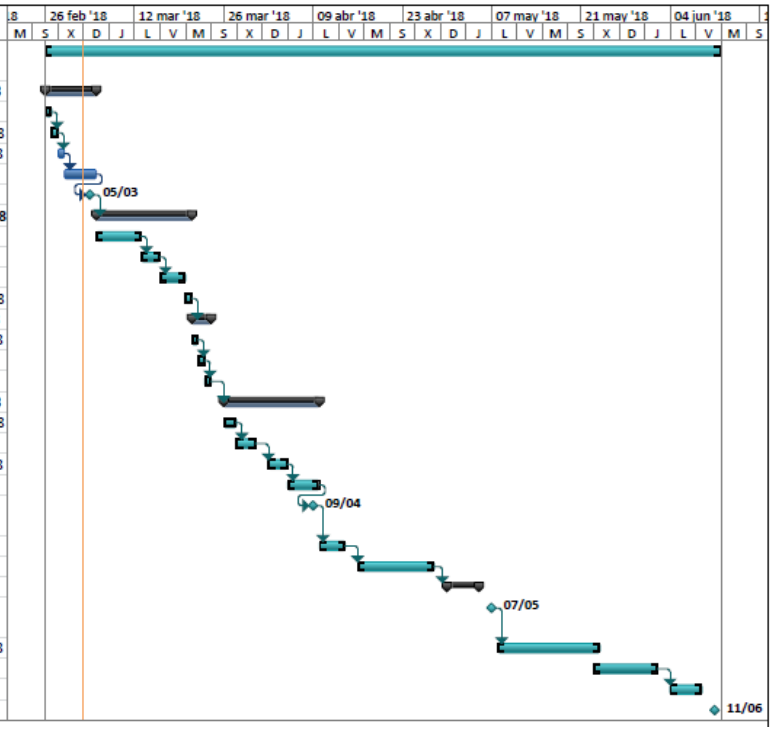


Ilustración 1. Diagrama de Gantt - Planificación