Sistema de Inteligencia de negocio para el análisis publicitario

Juan Manuel Heras Gomez
Master en Ingeniería Informática
M1.221. TFM – Business Intelligence

David Amorós Alcaraz
Maria Isabel Guitart Hormigo

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

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Sistema de Inteligencia de negocio para el análisis publicitario*
Nombre del autor:	*Juan Manuel Heras Gomez*
Nombre del consultor/a:	*David Amorós Alcaraz*
Nombre del PRA:	*María Isabel Guitart Hormigo*
Fecha de Entrega:	*01/2018*
Titulación o programa:	*Master en Ingeniería Informática*
Área del Trabajo Final:	*M1.221 – TFM-Business Intelligence*
Idioma del trabajo:	*Castellano*
Palabras clave:	*Business Intelligence Negocio*

Resumen del Trabajo:

El presente proyecto tiene como finalidad la implementación de un sistema Business Intelligence que facilite la toma de decisiones a partir de la explotación de los datos obtenidos de una campaña publicitaria llevada a cabo en distintos centros comerciales para determinar su efectividad.

A partir de una planificación y definición de tareas, se determinó que la metodología que más se ajustaba a este proyecto debido a su relación con el desarrollo se ha elegido una metodología ágil, más concretamente Scrum.

El principal objetivo del TFM es obtener unas conclusiones tras el análisis de los datos para alcanzar este fin se han generado una serie de entregables. Por un lado, la creación de un entorno de trabajo BI mediante la suite de herramientas de BI de Pentaho. Por otro lado, los distintos archivos relacionados con la explotación de los datos (ficheros ETL, cubos en XML y las queries y paneles de mando de saiku).

Las conclusiones obtenidas tras realizar este TFM han sido muy positivas. Para ser una primera toma de contacto en el ámbito del BI me ha permitido conocer todo el proceso de explotación y análisis de datos previos a la toma de decisiones.

Abstract:

This Project has the objective of train in the Business Intelligence area creating an environment simulating the real life. The project focus is to see the effectiveness of an advertising campaign and see the conclusions we can obtain from the data.

After done a planning where It's detailed every stage of the project It has been chosen Scrum as the methodology to work on.

The tool chosen for this BI project was the Pentaho suite. Pentaho provides all the tools needed for a complete BI project using ETL tools, Schema workbench and the BI.

In a short, this project is very interested and was a good first contact in the BI area.

Índice

1. Introducción 1
1.1 Contexto y justificación del trabajo 1
1.2 Objetivos del Trabajo 1
1.3 Enfoque y método seguido 2
1.5 Breve resumen de los productos obtenidos 5
1.6 Breve descripción de los capítulos de la memoria 5
2. Análisis de las herramientas del mercado 7
2.1 Análisis de las herramientas 7
2.2 Herramienta seleccionada: PENTAHO 13
3. Entorno de trabajo 14
3.1. Requisitos mínimos para instalar PENTAHO. 14
3.2. Instalación del Entorno. 15
3.3. Herramienta para el Análisis del Entorno BI. 16
4. Diseño e implementación del Datawarehouse 18
4.1. Diseño del DWH. 18
4.1.1 Cubo Relación Ventas-Impactos. 21
4.1.2. Cubo Visitas. 23
4.2. Implementación del DWH. 24
4.3. Fuentes de datos. 26
5. Diseño e implementación de los procesos ETL 28
5.1. Introducción 28
5.2. Diseño ETL. 30
5.3. Implementación del ETL. 31
5.3.1. Hechos_Impactos_Ventas 31
5.3.2. Hechos_Visitas 32
5.3.3. Dimensión Producto y Familia 33
5.3.4. Dimensión Shopping_Center 33
5.3.5. Dimensión Tiempo 34
6. Creación del Entorno BI para el análisis 35
6.1. Creación de los cubos 35
6.2. Diseño e implementación de los paneles de mando. 37
7. Análisis de la información 38
7.1. Análisis de los datos. 38
7.2. Queries Impactos-Ventas. 38
7.3. Queries Visitas. 45
7.4. Paneles de Mando 48
7.5. Conclusiones de la campaña publicitaria. 50
8. Conclusiones 51
9. Glosario 52
10. Bibliografía 53
11. Anexos 54
Anexo 1: Diagrama de Gantt 54
Anexo 2: Relación de archivos ETL generados 55
Anexo 3: 0_hechos_impactos_ventas_job.kjb 56
Anexo 3: 1_Hechos_Impactos_Ventas_Gran_Jorquera_etl.ktr 57
Anexo 4: 2_Hechos_Impactos_Ventas_Girocentre_etl.ktr 58

Anexo 5: 3_Hechos_Impactos_Ventas_la_maquinista_etl.ktr 59
Anexo 6: 4_Hechos_Impactos_Ventas_mercat_del_pla_etl.ktr 60
Anexo 7: 5_Hechos_Impactos_Ventas_les_gabarres_etl.ktr 61
Anexo 8: 6_Hechos_Impactos_Ventas_la_fira_etl.ktr 62
Anexo 9: Hechos_Visitas_etl.ktr 63
Anexo 10: producto_familia_etl_v2.ktr 64
Anexo 11: shopping_center_etl_v2.ktr 65
Anexo 12: cubo.xml 66
Anexo 13: Panel Ventas – Impactos 67

Lista de figuras

Figura1. Ejemplo de dashboard de BIRT. 7
Figura 2: Dashboard de SpagoBI 8
Figura 3.Dashboard de Pentaho 9
Figura 4: visualización de datos con Jaspersoft 11
Figura 5: Dashboard de Jaspersoft 11
Figura 6. Dashboard de Qlikview 12
Figura 7. Cubos 20
Figura 8. Detalle Diseño Cubo Relación Ventas-Impactos 21
Figura 9. Detalle Diseño Cubo Visitas 23
Figura 10. Tablas Datawarehouse 24
Figura 11. Tablas dim_shoppingcenter 24
Figura 12. Tablas dim_time 25
Figura 13. Tablas dim_producto_familia 25
Figura 14. Tabla hechos_impactos_ventas 25
Figura 15. Tabla hechos_visitas 25
Figura 16. SCHEMA DATACAMPAIGN 26
Figura 17. Ficheros de datos (CSV y Excel) 27
Figura 18. Area de Staging según kimball 29
Figura 19. Ejemplo ETL hechos_impactos_ventas_Gran_Jorquera 31
Figura 20. Job tabla de hechos impactos-ventas 32
Figura 21. ETL tabla de hechos_visitas 32
Figura 22. ETL dimensión producto y familia 33
Figura 23. ETL dimensión jerarquica shopping_center 33
Figura 24. ETL dimensión tiempo 34
Figura 25. Vista dimensiones Schema del cubo. 35
Figura 26. Cubo Ventas –Impactos y Cubo Visitas. 36
Figura 27. Efectividad de los impactos por zona. 39
Figura 28. Efectividad de los impactos por ciudad 39
Figura 29. Relación familias de productos, artículos impactados y ventas 41
Figura 30. Relación familias impactos y ventas 41
Figura 31. Relación familias impactos y ventas 42
Figura 32. Efectividad de los impactos por tiempo. 42
Figura 33. Gráfico Efectividad de los impactos por tiempo. 43
Figura 34. Efectividad zona familia. 44
Figura 35. Efectividad zona familia y producto. 44
Figura 36. Efectividad zona y tiempo. 45
Figura 37. Detalle Visitas. 45
Figura 38. Total Visitas por mes. 46
Figura 39. Vista visitas por día 47
Figura 40. Total Visitas por Zona 47
Figura 41. Panel Ventas-Impactos y Visitas 48
Figura 42. Panel Total Visitas – Total Ventas 49

1. Introducción

1.1 Contexto y justificación del trabajo

El marketing visual ha demostrado incrementar el número de clientes y su efecto sobre las ventas. Sus radicales funciones distintivas ayudan al diseño de tiendas de todo tipo de comercio a nivel global donde, los resultados demuestran, que el diseño de un establecimiento tanto interno como externo también surgen para atraer a los clientes. Sus elementos toman acción en las decisiones de los compradores de forma que inconscientemente se interesan por la facilidad de compra.

Las nuevas tecnologías han hecho evolucionar el concepto de Marketing Visual incorporando nuevas formas de captar la atención desde la reproducción de anuncios en todo tipo de pantallas interconnectadas, publicidad dentro de las aplicaciones o en las redes sociales más utilizadas.

Mediante el presente trabajo se pretende crear un entorno Business Intelligence (BI) que posibilite el análisis de la información generada durante una campaña publicitaria basada en la reproducción de pequeños anuncios en forma de impactos visuales en pantallas o monitores situados dentro de varios centros comerciales. El objetivo es poder valorar si la reproducción de estos pequeños anuncios de forma controlada hace que el articulo anunciado incremente el número de ventas respecto a otros artículos publicitados de forma tradicional. A su vez, la información suministrada podrá ser analizada en forma de cubos de datos lo que permitirá un análisis más exacto y la extracción de conclusiones mucho más elaboradas.

1.2 Objetivos del Trabajo

Tal y como se ha comentado en el punto anterior el principal objetivo del trabajo ***es poder valorar si la reproducción de estos pequeños anuncios de forma controlada hace que el articulo anunciado incremente el número de ventas respecto a otros artículos publicitados de forma tradicional***.

El objetivo de este trabajo es el diseño e implementación de un sistema Business Intelligence que facilite la adquisición, el almacenamiento y la explotación de datos obtenidos de la campaña publicitaria llevada a cabo en distintas ubicaciones.

Así pues, los objetivos específicos del trabajo son:

1. Diseñar un almacenamiento de datos (Data Warehouse) que permita almacenar la información adquirida de los distintos orígenes.

2. Implementar este almacen de datos y programar los procesos ETL (extracción, transformación y carga) que permita alimentar el Data Warehouse a partir de los ficheros base facilitados.

3. Analizar las diferentes plataformas BI Open Source disponibles en el mercado que nos permitan explotar la información almacenada.

4. Elegir e implantar una de estas soluciones Open Source de tal forma que se disponga de una capa de software para el análisis de la información.

1.3 Enfoque y método seguido

Tras analizar varias alternativas y estudiar los distintos enfoques y métodos para desarrollar un proyecto de BI, finalmente me he decantado por aplicar **SCRUM** al desarrollo del proyecto. Scrum es una metodología ágil de gestión de proyectos cuyo objetivo primordial es elevar al máximo la productividad de un equipo. Reduce al máximo la burocracia, y se centra en producir software que funcione y en ofrecer resultados visibles en poco tiempo, incluso ante requerimientos inestables. Sólo abarca prácticas de gestión sin entrar en las prácticas de desarrollo. Se caracteriza por: [1]

- Adoptar una estrategia de desarrollo incremental, en lugar de la planificación y ejecución completa del producto.
- Basar la calidad del resultado más en el conocimiento tácito de las personas en equipos auto organizados, que en la calidad de los procesos empleados.
- Solapamiento de las diferentes fases del desarrollo, en lugar de realizar una tras otra en un ciclo secuencial o en cascada.

[2] Esta metodología se emplea en entornos que trabajan con requisitos inestables y que requieren rapidez y flexibilidad, situaciones frecuentes en el desarrollo de software y en los proyectos de BI.

Los proyectos de Business Intelligence no son propiamente "proyectos de desarrollo de software". Las aplicaciones de BI son las que son, y los proyectos suelen consistir en analizar las necesidades de los usuarios, y configurar las aplicaciones y los informes para cubrir esas necesidades. En un proyecto de BI, no se crea software, se crean "cubos", "pantallas", "informes", "indicadores", "dimensiones"... y toda una manera nueva de trabajar con la información.

La idea básica es que el desarrollo se hace de manera cíclica e incremental. En cada "ciclo de desarrollo" (sprint), el "equipo de desarrollo" (scrum team) selecciona el subconjunto de funcionalidades (sprint backlog) que se incorporarán al producto, y durante el siguiente ciclo se centrarán únicamente en completar esas funcionalidades y entregárselas al cliente. Los equipos son pequeños (máximo 8 u 9 personas). Los ciclos de desarrollo son cortos (entre 3 y 5 semanas,

habitualmente), por lo que se tiene una buena visibilidad sobre el avance del proyecto.

Respecto a lo indicado en el párrafo anterior, es evidente que la metodología será aplicada parcialmente dado que yo seré el único integrante del equipo y las reuniones diarias serán simuladas por un estudio en voz alta de en qué estado me encuentro del proyecto y que voy a poder realizar durante los próximos días, generando un KanBan que me permita comprobar en qué estado se encuentra cada fase. Por otro, lado los ciclos de desarrollo serán de 3 a 10 días según la planificación que veremos en el siguiente punto.

Otro punto por el que he considerado utilizar SCRUM es debido a su orientación a entregables ya que consideraremos a las PACs como las fechas de entregas (en el siguiente punto se especifican que se entregará en cada PAC). Además, SCRUM permite el solapamiento de las fases, este punto es interesante a priori dada mi poca experiencia en el campo del Business Intelligence podría suponer que alguno de los hitos definidos en la planificación pudiera ser solapado por otro sin problemas, dado que me encuentro trabajando es posible que por motivos laborales tenga que solapar algunas fases del proyecto para cumplir con la planificación.

1.4 Planificación del Trabajo

Para la planificación del proyecto se han tenido en cuenta las fechas de entrega de las PACs, considerando estás como puntos de control para la realización de los entregables. Por otro lado, tal y como anteriormente se ha comentado por motivos laborales es posible que algunos días planificados para realizar tareas entre semana no sea posible por lo que se han considerado algunos fines de semana para poder cumplir con la planificación.

La planificación se ha dividido en 12 fases principales los cuales en algunos casos para facilitar su realización han sido divididos a su vez en varios hitos.

Nombre de tarea	Duración	Comienzo	Fin
Sistema de Inteligencia de negocio para el análisis publicitario	**76 días**	**lun 25/09/17**	**lun 08/01/18**
1. Planificación del proyecto	**6 días**	**lun 25/09/17**	**lun 02/10/17**
1.1. Preparar documentación proyecto	2 días	lun 25/09/17	mar 26/09/17
1.2. Crear planificación del proyecto	2 días	mié 27/09/17	jue 28/09/17
1.3. Revisión y provisión de recursos	1 día	vie 29/09/17	vie 29/09/17
1.4. Entregable 1. Planificación (PAC 1)	1 día	lun 02/10/17	lun 02/10/17
2. Análisis de las herramientas del mercado	**4 días**	**mar 03/10/17**	**vie 06/10/17**
2.1. Analisis de herramientas Open Source del mercado	3 días	mar 03/10/17	jue 05/10/17
2.2. Documentar Analisis en Memoria	1 día	vie 06/10/17	vie 06/10/17
3. Propuesta de la herramienta BI	**4 días**	**lun 09/10/17**	**jue 12/10/17**

3.1. Realizar propuesta técnica indicando pros y contras	3 días	lun 09/10/17	mié 11/10/17
3.2. Presentar propuesta técnica	1 día	jue 12/10/17	jue 12/10/17
4. Creación del entorno de trabajo	**5 días**	**vie 13/10/17**	**jue 19/10/17**
4.1. Revisar requisitos del Servidor para instalación del software BI	1 día	vie 13/10/17	vie 13/10/17
4.2. Instalación Servidor	3 días	lun 16/10/17	mié 18/10/17
4.3. Instalación Software BI	1 día	jue 19/10/17	jue 19/10/17
5. Diseño del DWH	**12 días**	**vie 20/10/17**	**lun 06/11/17**
5.1. Analizar necesidades del proyecto	9 días	vie 20/10/17	mié 01/11/17
5.2. Documentar diseño en Memoria	2 días	jue 02/11/17	vie 03/11/17
5.3. Entregable Documentación diseño DWH. (PAC2)	1 día	lun 06/11/17	lun 06/11/17
6. Implementación del DWH	**5 días**	**mar 07/11/17**	**lun 13/11/17**
7. Diseño de los procesos ETL	**8 días**	**mar 14/11/17**	**jue 23/11/17**
8. Implementación de los procesos ETL	**9 días**	**vie 24/11/17**	**mié 06/12/17**
8.1. Implementación de procesos	8 días	vie 24/11/17	mar 05/12/17
8.1. Entregable Documentación estado proyecto. (PAC3)	1 día	mié 06/12/17	mié 06/12/17
9. Creación del entorno BI para el análisis	**11 días**	**jue 07/12/17**	**jue 21/12/17**
10. Análisis de los datos	**7 días**	**vie 22/12/17**	**lun 01/01/18**
11. Conclusiones	**4 días**	**mar 02/01/18**	**vie 05/01/18**
12. Entrega del Proyecto	**1 día**	**lun 08/01/18**	**lun 08/01/18**

Como podemos observar en verde se destacan los cuatro entregables planificados en las fechas de las PACs indicadas en la planificación de desarrollo del proyecto.

Para el primer entregable se realizará la planificación del proyecto. Este hito está planificado para el día 02/10/2017. Con este primer entregable se entregará una primera versión del documento TFM que incluirá la introducción explicando en que consiste el proyecto y el propio documento, el enfoque y metodología elegida y la planificación detallada del proyecto.

En el segundo entregable, es el que consta de más fases, aunque la previsión es que no sean demasiado extensas a excepción de la fase de diseño. Para este entregable se habrá creado el entorno de trabajo con la herramienta BI elegida y se preparará el Diseño del DWH. Está planificado para el día 06/11/2017 y tiene como objetivo entregar una nueva versión del documento del TFM. Para este entregable se espera tener completadas las fases 2,3,4 y 5 del proyecto y en el documento TFM se crearán un capítulo para cada una de las fases explicando cómo han sido realizadas.

El tercer entregable, se prevé más difícil dada mi poca experiencia, ya que lo dedicaremos a la implementación del DWH, y al diseño e implementación de los procesos ETL, fases 6, 7 y 8 de la planificación. Se entregará una nueva versión del TFM incluyendo los capítulos de

estas fases documentadas. La fecha prevista de entrega es el 06/12/2017.

El cuarto y último entregable será el propio TFM. Este último entregable incluirá la Creación del Entorno de BI para el análisis, el análisis de los datos y las Conclusiones. Se entregará la última versión del TFM junto con una presentación del mismo para su calificación por parte del tribunal. La fecha prevista de entrega es el 08/01/2018.

Para ver el detalle de la planificación mediante un Diagrama de Gantt consultar el Anexo 1 de este documento.

Con respecto a los costes me he decantado por realizar una valoración por jornada laboral. Dada la experiencia en la materia se estima un coste por jornada de 90€/día. De modo que los **costes estimados** del proyecto serán de 76 días x 90€/día = **6840 €**

1.5 Breve resumen de los productos obtenidos

El principal producto obtendremos al finalizar este proyecto será una análisis de los datos que nos permitirá la toma de decisiones con respecto a la realización de una campaña publicitaria.
No obstante, en el proceso se generarán entregables los cuales están en línea con los entregables estipulados en la planificación. En primer lugar, se creará un entorno de trabajo provisto de una herramienta BI y un datawarehouse. En segundo lugar, se generarán archivos ETL para la transformación de los datos proporcionados. En tercer lugar, se generarán archivos XML tras la creación de los cubos. Y por último, se generarán una serie de consultas y paneles de mando a través de la herramienta BI.

1.6 Breve descripción de los capítulos de la memoria

Los capítulos de la memoria, también han sido creados en línea con la estructura de la planificación del proyecto.
El primer capítulo, es la introducción a este TFM, donde se introduce el problema planteado y se describe el documento de la memoria.

El segundo capítulo, será el Análisis de las herramientas del mercado cuya finalidad es tras estudiar las distintas herramientas del mercado seleccionar la que más se adecue al proyecto.

El tercer capítulo, Entorno de trabajo explicará las requerimientos y características del entorno en el que se va a instalar la herramienta BI seleccionada y desarrollar el proyecto.

El cuarto capítulo, Diseño e implementación del Datawarehouse, pretende explicar cómo se ha efectuado el proceso de diseño, como se

ha decidido hacerlo y como ha sido su implementación mostrando las distintas alternativas.

El quinto capítulo, Diseño e implementación de los procesos ETL, explicarán cómo se han diseñado e implementado los distintos procesos ETL generados para la carga de datos en el DWH.

El sexto capítulo, Creación del Entorno BI para el análisis, explica cómo se han generado los cubos y como se han diseñado y creado los paneles de mando.

El séptimo capítulo, Analisis de la Información, tiene como intención mostrar las distintas queries y paneles de mando realizados para el análisis de la información y las conclusiones de la campaña publicitaria.

El octavo capítulo, Conclusiones, se explican las conclusiones del TFM dando una opinión personal de cómo ha sido.

Los capítulos Noveno, décimo y undécimo, son el glosario, bibliografía y anexo respectivamente.

2. Análisis de las herramientas del mercado

2.1 Análisis de las herramientas

La oferta actual de herramientas de BI es muy grande y podría ocupar la mayor parte del proyecto por lo que para el análisis de las herramientas me he centrado en las Open Source y versión Free Edition más utilizadas y mejor valoradas del mercado.

A continuación, se explican las herramientas analizadas [3]:

- **BIRT:** es un programa de BI de código abierto que "a menudo se considera el estándar de la industria". Esta percepción tiene mucho de verdad, ya que BIRT cuenta con "más de 12 millones de descargas y más de 2,5 millones de desarrolladores en 157 países". Sus usuarios incluyen pesos pesados como Cisco, S1 e IBM (que también es un patrocinador de BIRT). También tienen madurez para ellos, ya que han estado presentes desde la (segunda) administración de Bush.

Figura1. Ejemplo de dashboard de BIRT.

BIRT puede crear una variedad de informes, desde documentos textuales hasta tablas de referencias cruzadas y gráficos de barras estándar. Junto con estos conceptos básicos de BI, BIRT también puede abordar tareas ligeramente más avanzadas, "como agrupar en sumas, porcentajes de totales generales y más". BIRT también puede integrarse con otras aplicaciones. BIRT también es excelente si ejecuta una aplicación que utiliza Java, ya que componentes como el motor de informes y el motor de creación de gráficos de BIRT se integran fácilmente con aplicaciones y programas que usan Java.

PROS:

- Se integra fácilmente con otras aplicaciones.
- Considerada como un estándar
- Gran Variedad de Informes

CONTRAS:

- Requiere gran conocimiento de código.
- Sólo módulo BI. No dispone de módulo para ETL, por ejemplo.

- **SpagoBI:** Desde la última actualización, SpagoBI se ha combinado con Knowage. SpagoBI todavía está disponible, y sigue siendo de código abierto, pero ahora se ofrece junto con Knowage y Knowage Community Edition. Spago es, de acuerdo con su sitio, "la única suite de Inteligencia empresarial de código abierto".

  También son lo suficientemente competentes como para haber ayudado recientemente a una importante empresa de logística de Colombia, así como a varias ciudades holandesas que deseaban administrar los gastos de atención médica.

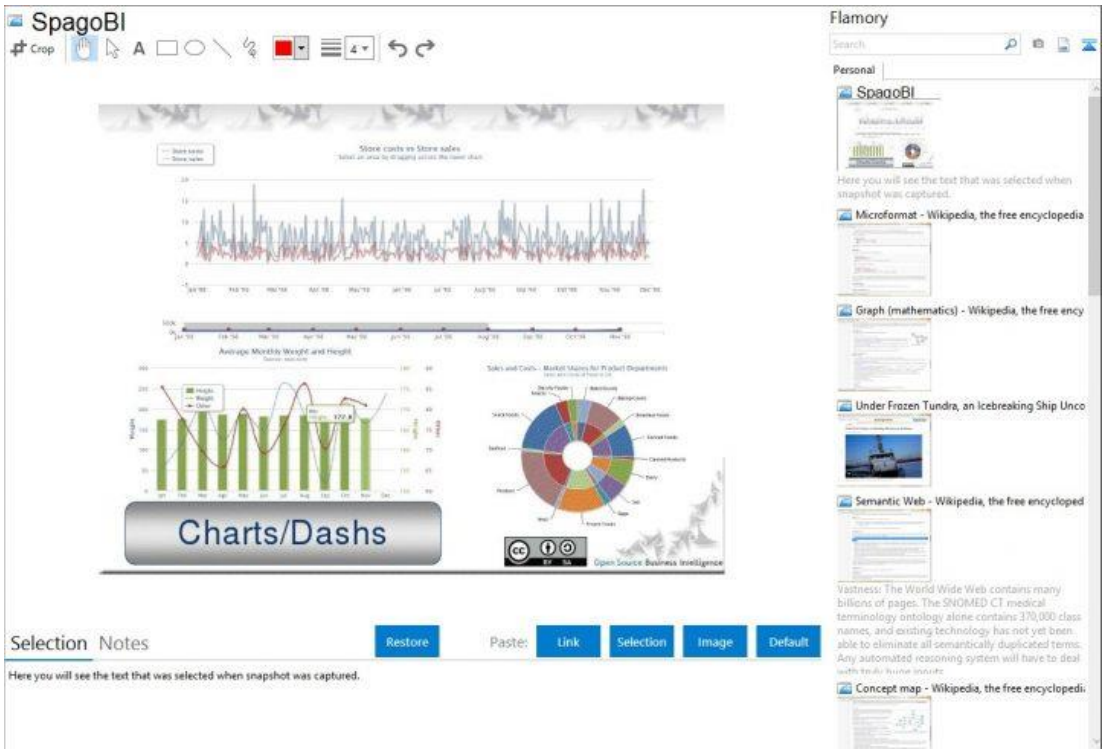


Figura 2: Dashboard de SpagoBI

[4] Es una multiplataforma ya que cubre y satisface todos los requisitos de BI (Business Intelligence), tanto en términos de análisis y de gestión de datos, administración y seguridad. En el mundo analítico ofrece soluciones para la presentación de informes, análisis multidimensional (OLAP), minería de datos (Data Mining), tableros de mando (Dashboard) y consultas ad-hoc. Añade módulos originales para la gestión de procesos de colaboración a través de análisis dossiers y el análisis de geo-referencia.
Cuenta con herramientas para la extracción de datos, transformación y carga (ETL y apoya al administrador en el mantenimiento de los

documentos analíticos, la gestión para el control de versiones y la aprobación del flujos de trabajo (workflow).

PROS:

- Dispone de varias herramientas BI tanto para análisis como para gestión de datos.
- La suite es totalmente Open Source.

CONTRAS:

- SpagoBI es un poco difícil de configurar y usar cuando eres un nuevo usuario.

- **Pentaho Community:** Pentaho fue comprado por Hitachi, y ahora constituye una parte de Hitachi Ventara. Sin embargo, todavía se le llama Pentaho. Ofrecen una gran cantidad de elementos básicos de BI útiles, como visualizaciones desde gráficos de calor hasta el mapeo geográfico, así como informes en productos tan diversos como PDF, Excel, HTML y otros. Ofrecen una Edición Enterprise (para la cual tendrás que comprar una suscripción) y una Community Edition (gratuita y de código abierto). Al igual que con Jaspersoft, la empresa necesitará a alguien que sepa cómo codificar si realmente desea aprovecharla.

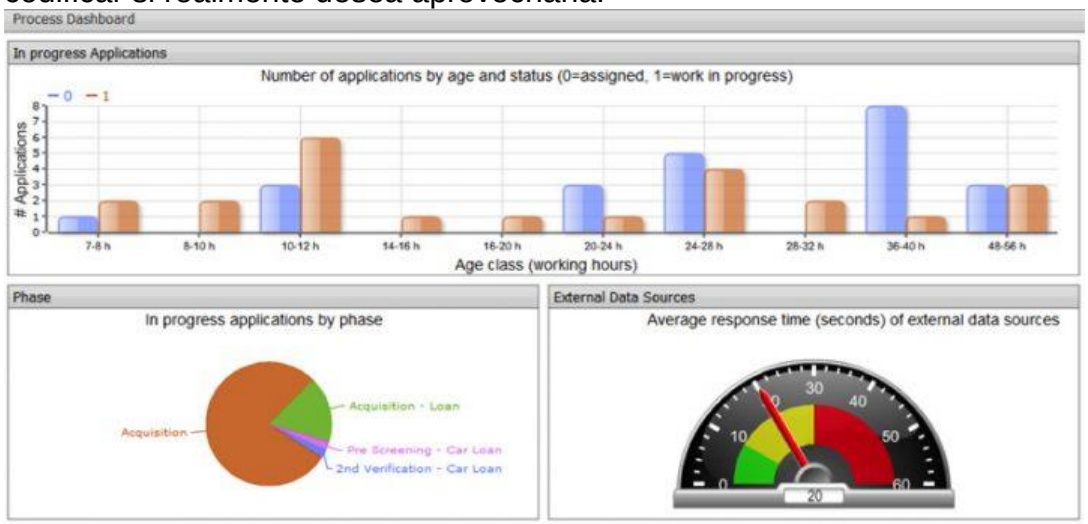


Figura 3.Dashboard de Pentaho

Pentaho definitivamente trata de tomarse en serio la idea de una comunidad de desarrolladores. Hay una lista de grupos de la vida real, con enlaces a meetup.com, pero muchos de los grupos no han tenido reuniones en meses. La excepción es Business Intelligence Big Data Integration ETL Analysts. Están basados en Nueva York, pero dos de sus últimas tres reuniones fueron seminarios web gratuitos, por lo que la geografía no es un problema.

[5] En resumen, Pentaho se define a sí mismo como una plataforma de BI “orientada a la solución” y “centrada en procesos” que incluye todos los principales componentes requeridos para implementar soluciones basadas en procesos tal como ha sido concebido desde el principio.

Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e

informes integrados con un motor de workflow de procesos de negocio. La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de procesos, actividades, además es capaz de presentar y entregar la información adecuada en el momento adecuado.

PROS:

- Módulos de ETL, Integración de datos, análisis y framework.
- Gran comunidad de desarrolladores
- No tiene fecha de caducidad, es decir, la versión community no tiene tiempo límite de uso.
- Fácil manejo
- La herramienta de integración de datos es muy fuerte. Es fácil de usar y muy accesible para los usuarios de negocios.
- El componente de integración de datos proporciona la capacidad de conectarse a casi cualquier fuente de datos como entrada y salida, de manera rápida y fácil
- El componente de integración de datos permite una vista previa rápida de los datos en un formato de tipo de vista dinámica que permite el análisis de datos en una etapa temprana de desarrollo

CONTRAS:

- Requiere especialización para la realización de determinadas acciones como su integración con otras herramientas.
- La interfaz del servidor de BI no parece ser muy atractiva para los usuarios finales. Las críticas indican que los usuarios finales desean una pantalla más bonita y una navegación más sencilla.
- El workbench del Schema no es fácil de desarrollar.

- **Jaspersoft Community:** Jaspersoft es una de las herramientas de código abierto más populares del mercado. Fue comprada por Tibco en 2014 por 185 millones de dólares, lo que demuestra su calidad y su dilatada experiencia.

  Jaspersoft Community es la oferta gratuita de la compañía la cual ofrece cinco de sus programas en ediciones comunitarias:

  - **Servidor JasperReports:** un programa para diseñar informes que pueden ser integrados o utilizados por sí mismos
  - **JasperReports Library:** una biblioteca de Java que le permite usar datos de cualquier fuente y exportar informes en HTML, PDF, Excel y otros formatos
  - **Jaspersoft ETL:** un motor de integración de datos que transforma los datos de su negocio en información que puede usarse en un informe.
  - **Jaspersoft Studio:** un programa que le permite diseñar y personalizar los informes que incrustará en JasperReports Server.
  - **Visualize.JS:** un marco de JavaScript que incorpora aplicaciones de Jaspersoft en cualquier software que ya use.

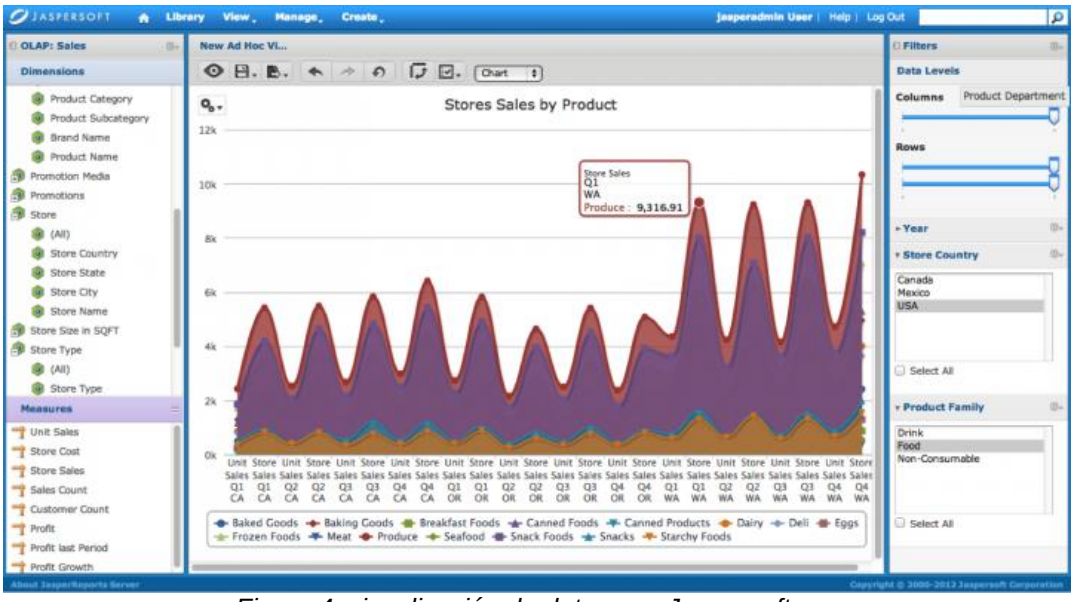


Figura 4: visualización de datos con Jaspersoft

El sitio web de Jaspersoft afirma que el 94% de los clientes recomiendan su producto y el 90% recomienda su apoyo. Aunque ese hecho es de su sitio web, no se puede discutir con su impresionante lista de clientes: todos, desde Groupon a Time Warner Cable, al gobierno de Columbia Británica y la Universidad de Vanderbilt.

Al igual que muchos programas de código abierto, Jaspersoft tiene una comunidad de desarrolladores, un foro en línea para las personas que usan y, bueno, desarrollan ese código fuente abierto en un programa más completo.

Figura 5: Dashboard de Jaspersoft

Lo mejor de las comunidades de desarrolladores para cualquier programa de código abierto es que tienes una red de soporte potencial, o al menos personas con inquietudes similares. La desventaja es que, dado que las comunidades de desarrolladores están pobladas por usuarios, en lugar de asesores de servicio al cliente pagados, se arriesga cuando busca ayuda.

PROS:

- Varios programas en su versión Communitiy para ETL, Diseño, Framework y Librería.
- Extensa comunidad de desarrolladores.

CONTRAS:

- Acceso ilimitado gratuito durante 60 días
- Requiere especialización

- **<u>Qlikview Personal Edition:</u>** Qlik, uno de los tres únicos líderes en el cuadrante mágico de Business Intelligence de Gartner 2017, ofrece versiones gratuitas y "personales" de tres de sus programas: QlikView Personal, Qlik Sense Desktop y Qlik Sense Cloud.

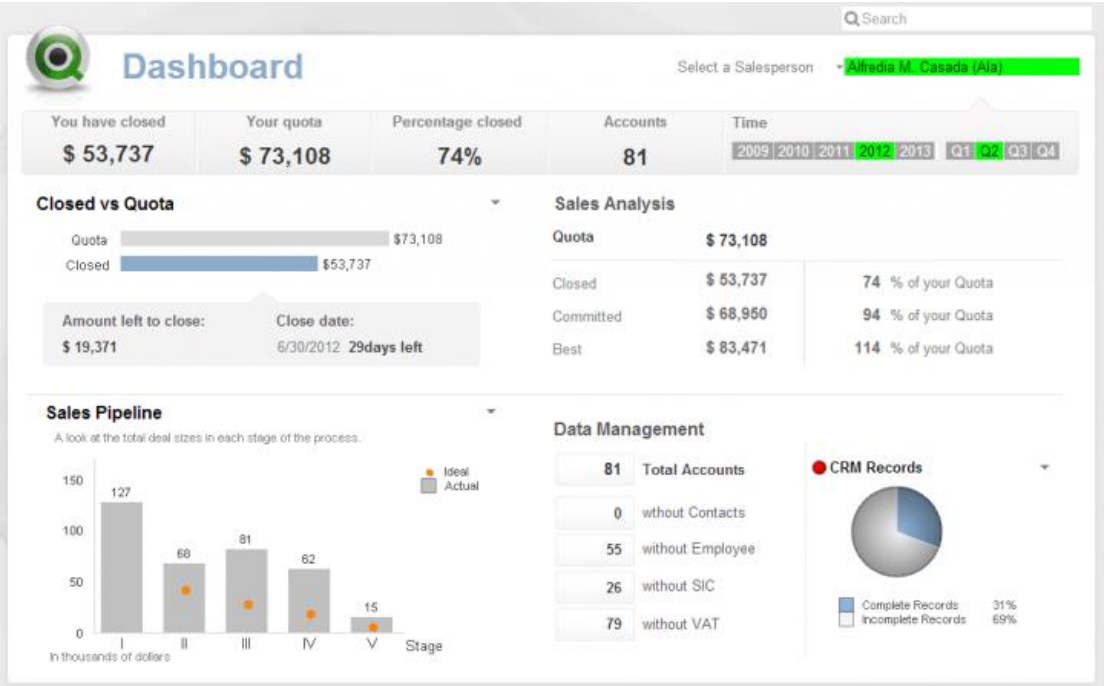


Figura 6. Dashboard de Qlikview

Las ediciones personales de Qlik son las mejores para usar en solitario. Tienen las mismas funciones que la versión completa y no están programadas (como en pruebas de 30 días), así puedes obtener un producto sólido. No ofrecen la misma capacidad para colaborar y compartir datos que obtendría con las versiones completas, ni dispone de módulos para ETL por ejemplo.
Se trata de una herramienta ideal para crear tus cuadros de mando a partir de ficheros Excel o conectando directamente a las BBDD pero explotando dichos datos directamente.

PROS:

- Sencillo de usar. No se necesita apenas conocimiento alguno para crear un cuadro de mando básico.
- Cuenta con una versión gratuita privada, perfecta para valorar su capacidad o emplear en una pequeña organización.
- Precio comparativamente inferior respecto a los principales competidores.
- Filtrado y navegación muy intuitiva a la que los usuarios se adaptan inmediatamente.
- Capacidades de desarrollo muy avanzadas para usuarios con conocimiento técnicos.

CONTRAS:

- Se trata de una herramienta de pago. Muy limitada en su versión Personal.

2.2 Herramienta seleccionada: PENTAHO.

Tras el análisis realizado en el punto anterior, se decide finalmente que la herramienta elegida sea **Pentaho**.

La elección de Pentaho es debido a que se trata de una herramienta contrastada en el mercado con más de 10 años de desarrollo a sus espaldas, lo cual nos permitirá encontrar documentación fácilmente y el adquirir conocimiento de la misma nos puede servir para aplicarlo en el ámbito profesional. Además cuenta una comunidad de desarrolladores bastante extendida.

Por otro lado, nos encontramos con una herramienta que proporciona una experiencia completa de aprendizaje de los conceptos del Business Intelligence dado que la suite Pentaho provee los distintos módulos de Business Intelligence, ETL, Reports y workbench schema en versión Community Edition. Esto permitirá una asimilación y un aprendizaje más completo de cómo funcionan las herramientas BI.

Otro punto a considerar, es que según las críticas, Pentaho es una herramienta fácil de utilizar lo que hará que la curva de aprendizaje sea más rápida.

Por último comentar que la versión Community Edition de Pentaho, a pesar de haber sido adquirida por Hitachi, se trata de una herramienta Open Source multiplataforma la cual podemos instalar en Linux sin problemas, lo que evitará tener que adquirir licencias para su uso en este proyecto.

3. Entorno de trabajo

3.1. Requisitos mínimos para instalar PENTAHO.

Pentaho tiene como objetivo dar cabida a diversos entornos informáticos. Esta lista proporciona detalles sobre los componentes de entorno y las versiones que admite. El software de servidor Pentaho es independiente del hardware y se ejecuta en equipos servidor que cumplen con estas especificaciones para el hardware mínimo y los sistemas de operación requeridos. Tal y como podemos encontrar en la documentación oficial de Pentaho los requisitos mínimos para su instalación son los siguientes: [6](subrayado en Verde se destacan los componentes a instalar en el Entorno de trabajo)

Server:

- **Pentaho Business Analysis Server**
- **Pentaho Data Integration Server**

Hardware 64 bits	Sistema Operativo 64 bits
Processor: • Apple Macintosh Pro Quad-Core or Macintosh Mini Quad-Core • Intel EM64T or AMD64 Dual-Core **RAM:** 8 GB with 4 GB dedicated to Pentaho servers **Disk Space:** 20 GB free after installation	• Apple Macintosh OS X Server 10.8 & 10.9 • CentOS Linux 5 & 6 • Microsoft Windows 2008 Server R2, 2012 Server R2 • Red Hat Enterprise Linux 5, 6 & 7 • Solaris 10 & 11 • Ubuntu Server 12.x, 14.x & 16.x

Workstation:

- Pentaho Aggregation Designer
- **Pentaho Data Integration**
- Pentaho Metadata Editor
- **Pentaho Report Designer**
- **Pentaho Schema Workbench**
- Pentaho Mobile

Hardware 32 o 64 bits	Sistema Operativo 32 o 64 bits
Processors: • Apple Macintosh Dual-Core, 64 bit only • Intel EM64T or AMD64 Dual-Core **RAM:** 2 GB RAM for most of the design tools, PDI requires 2 GB dedicated **Disk Space:** 2 GB free after installation **Minimum Screen Size:** 1280 x 960	• Apple Macintosh OS 10.8 & 10.9, 64 bit only • Microsoft Windows 7, 8 & 10 • Ubuntu Desktop 12.x, 14.x & 16.x • iPad 2, 3, & Air (iOS 6.x & 7.x)

3.2. Instalación del Entorno.

Para continuar con la filosofía Open Source como sistema se ha optado por el montaje de una máquina virtual mediante Virtual Box. Las ventajas que proporciona trabajar con una máquina virtual son muchas, desde poder realizar Snapshots previos a los cambios que realicemos sobre la misma o la versatilidad que brinda el poder copiar la máquina virtual a distintos equipos para de esta forma poder trabajar con ella en cualquier momento.

El sistema operativo elegido es Ubuntu Linux. La versión de Ubuntu elegida es la 16.04. El motivo de la elección de este sistema es debido a que dado que Pentaho permite su instalación en sistemas linux esto evitará, como anteriormente se ha comentado, la necesidad de tener que comprar licencias o bien de Sistema Operativo o bien de la propia herramienta. Por otro lado, el motivo de la elección de la distribución de Linux elegida ha sido Ubuntu debido a su fácil manejo y su extensa comunidad.

En lo que respecta al Datawarehouse se ha optado por instalar una base de datos MySQL con un SCHEMA dedicado. En dicho Schema se crearán las distintas tablas correspondientes a las Dimensiones decididas en el diseño del mismo a modo de OLAP.

Además, se creará base de datos para cargar los datos proporcionados en los ficheros Excel sobre la misma para simular un entorno OLTP (o base de datos relacional) y de esta forma simular al máximo lo que nos encontraríamos en un entorno real.

Dadas las limitaciones que se disponen para cumplir con los requisitos mínimos descritos en el punto anterior se ha creado una máquina virtual que se acerca en la medida de lo posible a lo que el fabricante demanda para un buen funcionamiento de la herramienta. En resumen, se ha creado una VM Ubuntu Linux con las siguientes características:

- RAM: 6GB
- vCPUs: 1
- Almacenamiento: 20 GB
- Interfaces de Red: 1 con configuración NAT
- Sistema Operativo: Ubuntu Linux 16.04
- Base de datos: MySQL
- DWH: Schema de MySQL
- Herramienta BI: Pentaho versión 6.1 con los siguientes módulos:
    - Pentaho Business Analytics Server o BI server
    - Pentaho Schema Workbench
    - Pentaho Report Designer
    - Pentaho Data Integration (antiguo kettle)

Instalación de Pentaho en Linux:

[7] Para la instalación de Pentaho he sigo los pasos indicados en la documentación oficial y en algunos foros. La instalación básicamente se basa en la descarga de los componentes indicados en el punto anterior y su descompresión. En mi caso, he guardado todos los componentes en una misma carpeta para facilitar su gestión.

Para que estos componentes funcionen hay una serie de dependencias. Según la documentación oficial es necesario instalar openjdk-8-jre, mysql server y el jdbc driver.

Para iniciar el entorno o el componente únicamente necesitaremos ejecutar el script de inicio en concreto que deseamos utilizar y esperar un tiempo a su arranque.

3.3. Herramienta para el Análisis del Entorno BI.

Para el análisis de la información a través de la plataforma Pentaho BI Server se ha decidido utilizar la herramienta **SAIKU**.

Saiku es un excelente visor OLAP que proporciona al usuario final una magnifica herramienta para realizar análisis de forma fácil e intuitiva.

Pero Saiku no es sólo eso. Es un buen ejemplo de cómo un proyecto Open Source puede ofrecer soluciones de excelente calidad a la vanguardia de la tecnología y delicada experiencia de usuario.

Saiku puede ser utilizado sólo si sólo quieres realizar análisis OLAP. Es un servidor independiente. En nuestro caso, puedes embeberlo en un servidor Pentaho como un pluguin de forma fácil y sencilla. Es un plugin de Pentaho, el cual podemos instalar desde el Marketplace de la herramienta.

Puedes utilizarlo como origen de datos. Es un backend. Y puedes, por ejemplo construir tu propia interfaz de usuario como han hecho en Inovia.fr . Que han hecho una interfaz PHP

La que más destaca de la herramienta es su interfaz fácil limpia e intuitiva que permite a los usuarios construir sus propias vistas arrastrando y soltando campos. Por otro lado, podemos analizar los datos en forma de tabla pivotante o en modo gráfico y podemos visualizar los datos de origen.

4. Diseño e implementación del Datawarehouse

4.1. Diseño del DWH.

Una fase muy importante de un proyecto de BI es el diseño del Datawarehouse. [8] Un data warehouse se encarga de extraer datos de las bases de datos operacionales o fuentes externas, transformar, consolidar, integrar, comprobar la integridad y centralizar los datos que la empresa genera en su actividad diaria de negocios y/o información externa con la que esté relacionada. Por otro lado, un data warehouse permite el acceso y manipulación de la información, a través de análisis multivariables, con el objetivo final de dar soporte al proceso de toma de decisiones.

Las particularidades de los data warehouse hacen, que los modelos de datos y las estrategias de diseño utilizadas para bases de datos operacionales, generalmente no sirvan para el diseño de un data warehouse. Esto implica, que se utilicen nuevas técnicas y estrategias de diseño.

Los data warehouse pueden crecer con el tiempo lo que puede conducir a la necesidad de dividir la información. Esta división normalmente se trata deuna división departamental y recibe el nombre de datamart. Tienen una estructura óptima, para analizar los datos desde el punto de vista de un departamento.

Este tipo de modelo de datos consta principalmente de dos tipos:

- **DIMENSIONES:** Representan factores por lo que se analiza un determinado área del negocio. Son pequeñas y usualmente están desnormalizadas. [9] En la construcción de cubos OLAP, las tablas de dimensiones son elementos que contienen atributos (o campos) que se utilizan para restringir y agrupar los datos almacenados en una tabla de hechos cuando se realizan consultas sobre dicho datos en un entorno de almacén de datos o data mart.
  Estos datos sobre dimensiones son parámetros de los que dependen otros datos que serán objeto de estudio y análisis y que están contenidos en la tabla de hechos. Las tablas de dimensiones ayudan a realizar ese estudio/análisis aportando información sobre los datos de la tabla de hechos, por lo que puede decirse que en un cubo OLAP, la tabla de hechos contiene los datos de interés y las tablas de dimensiones contienen metadatos sobre dichos hechos.

- **HECHOS:** Son el objeto de los análisis y están relacionados con las dimensiones. Son tablas muy grandes y suelen estar desnormalizadas. Se a menudo incluyen diferentes agregaciones como máximo, mínimo, media,… [10]En las bases de datos, y más concretamente en un data warehouse, una tabla de hechos (o tabla fact) es la tabla central de un esquema dimensional (en estrella o en copo de nieve) y contiene los valores de las medidas de negocio o dicho de otra forma los indicadores de negocio. Cada medida se toma mediante la intersección de las

dimensiones que la definen, dichas dimensiones estarán reflejadas en sus correspondientes tablas de dimensiones que rodearán la tabla de hechos y estarán relacionadas con ella.

Antes de explicar el diseño del DWH cabe recordar las preguntas que se quieren resolver en el proyecto. Esto nos ayudará a determinar qué datos necesitamos en los cubos y su creación. El tener claro las preguntas a responder nos permitirá eliminar ruido o realizar minería de datos para adaptar la información a las necesidades del proyecto. Las cuestiones a resolver son las siguientes:

- *En general la publicidad basada en impactos visuales, ¿es efectiva?*
- *¿Hay determinadas zonas geográficas donde los impactos son más efectivos?*
- *Según la familia de producto, ¿existe alguna relación entre los artículos impactados y el resultado obtenido?*
- *¿Son más efectivos los impactos según el momento temporal en que se producen?*
- *¿Existe algún artículo o familia que según la zona, su publicidad tiene mayor impacto? ¿Y según el momento en que se realiza?*

Además partimos de unos conocimientos previos de los datos que vamos a utilizar para responder a estas cuestiones. Esta información es la siguiente:

Tiempo:

- Se quiere agrupar toda la información con una granularidad mínima de día. Además se desea realizar agrupaciones semanales, mensuales, trimestrales y anuales.

Centros Comerciales:

- La campaña se llevará a cabo en distintos centros comerciales. Estos centros estarán clasificados como urbanos o extrarradio según se encuentren en el área metropolitana de Barcelona u otras comarcas.
- Se dispondrá de información sobre la ciudad y la zona donde se encuentra el centro comercial.
- También se dispondrá de información sobre el punto final donde se hacen los impactos. Aquí tendremos opciones de grandes pantallas del centro comercial, pantallas en las tiendas o monitores situados en los puntos de información.

Artículos:

- La campaña se realizará para distintas familias de productos. No todos los productos de una misma familia tendrán el mismo número de impactos visuales.

Indicadores:

- Para un determinado artículo, un punto publicitario tendremos los impactos que se han producido a lo largo del tiempo.
- Además también dispondremos de las ventas realizadas de aquel artículo durante aquel día.

- Paralelamente también tendremos el número de visitantes del centro comercial.

A partir de la información y las cuestiones planteadas, para el problema planteado en este proyecto se ha optado por un Modelo de datos en **Copo de Nieve**. [11] En las bases de datos utilizadas en (data warehousing), un esquema en copo de nieve es una estructura algo más compleja que el esquema en estrella. Se da cuando alguna de las dimensiones se implementa con más de una tabla de datos. La finalidad es normalizar las tablas y así reducir el espacio de almacenamiento al eliminar la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre las tablas (JOINS) lo que tiene un impacto directo sobre el rendimiento.

En las aplicaciones OLAP implementadas sobre bases de datos relacionales (ROLAP), un elemento clave es el Cubo OLAP. Estos cubos (también llamados hipercubos) almacenan grandes volúmenes de datos que posteriormente deben ser analizados en función de unos determinados parámetros. Al diseñar las tablas en las que se han de almacenar estos datos y parámetros, si se aplican las técnicas de normalización de bases de datos para optimizar el espacio requerido para guardar estos datos eliminando las redundancias, es habitual que se termine obteniendo un esquema en copo de nieve; en este tipo de esquemas se tiene una tabla central de hechos en la que se guardan las medidas del negocio que se quiere analizar, y en las tablas adyacentes se tendrán las dimensiones (parámetros) de que dependen los datos del negocio. Si por alguna dimensión se requiere más de una tabla se dice que el esquema resultante es un esquema en copo de nieve.

Para el problema planteado, y tras analizar las cuestiones se ha decidido la creación de dos cubos:

- **Cubo relación ventas-impactos** con las dimensiones Centro Comercial, Producto y Tiempo.
- **Cubo visitas** con las dimensiones Centro Comercial y Tiempo.

Figura 7. Cubos

A continuación, se explica con más detalle el diseño de estos cubos.

4.1.1 Cubo Relación Ventas-Impactos.

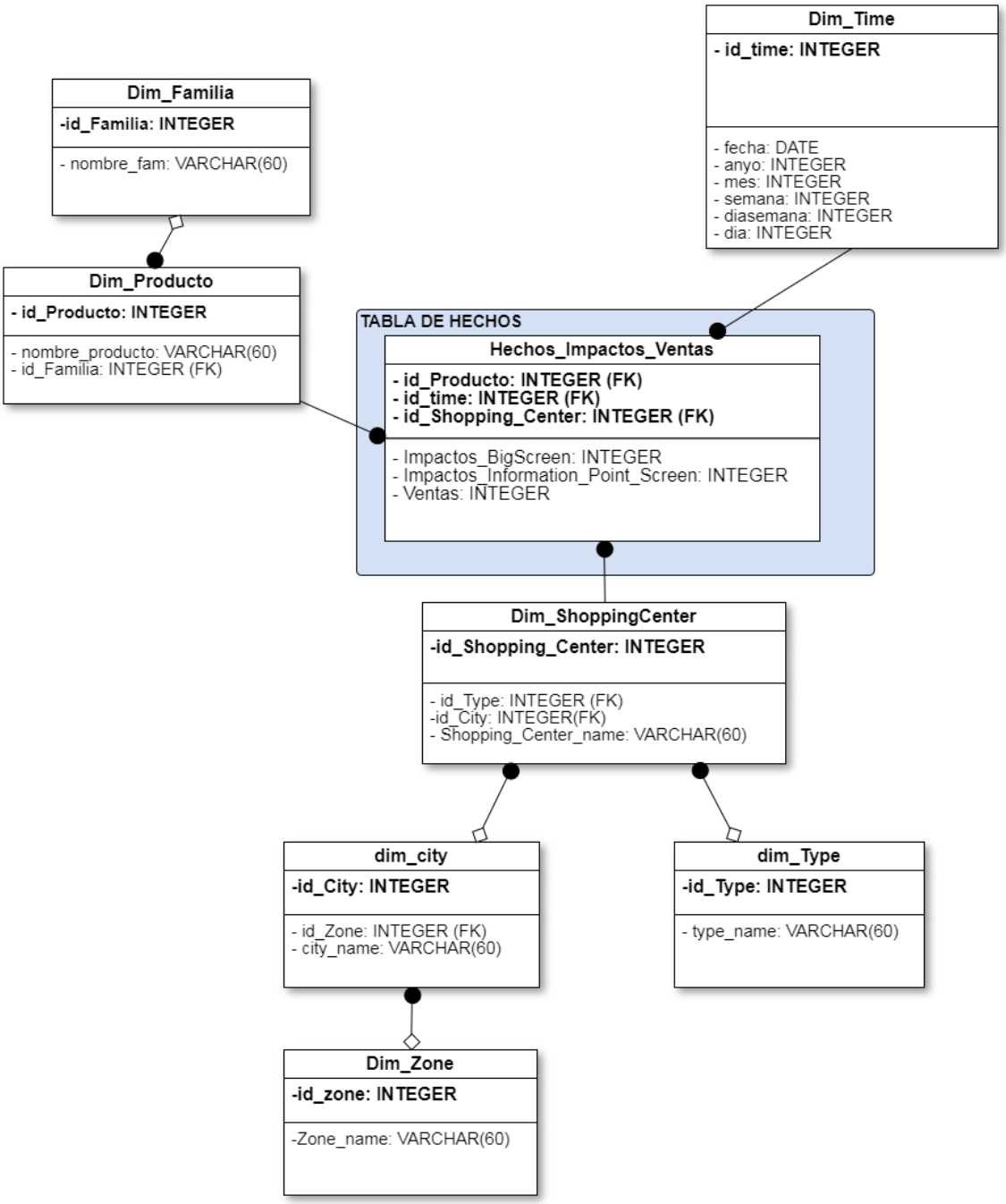


Figura 8. Detalle Diseño Cubo Relación Ventas-Impactos

Tal y como podemos observar en la imagen el cubo presenta una tabla de hechos con la relación de impactos y ventas. A partir de esta tabla de hechos se definen las dimensiones Producto, Tiempo, Zona y Punto de Impacto.

La tabla ***Hechos_Impactos_ventas*** está compuesta de la siguiente forma:
Las claves primarias indicarán el producto, el tiempo y el centro comercial en concreto:

- ***Id_Producto***: Clave ajena de tipo entero que hace referencia al producto en concreto del cual podemos ver el detalle en la dimensión producto.

- ***Id_Shopping_Center***: Clave ajena de tipo entero que hace referencia al centro comercial en concreto del impacto o venta del cual podemos ver el detalle en la dimensión ShoppingCenter.
- ***Id_Time***: Clave ajena de tipo entero que hace referencia al punto de impacto en concreto del cual podemos ver el detalle en la dimensión ShoppingCenter.

Dadas las claves indicadas anterirormente se mostrarán unos indicadores. A continuación se muestran los indicadores:

- ***Impactos_BigScreen***: Indicador de número de impactos realizados en grandes pantallas para ese punto del cubo en concreto.
- ***Impactos_Information_Point_Screen***: Indicador de número de impactos realizados en puntos de información para ese punto del cubo en concreto.
- ***Ventas***: Para ese punto del cubo en cocreto.

Las dimensiones creadas tal y como anteriormente se han comentado son Tiempo – Producto – Centro Comercial.

La dimensión Tiempo contará con una tabla cuya clave primaria será su idTiempo y con unos indicadores de fecha, año, trimestre, mes, semana y diasemana dado que se requiere una granularidad mínima.

La dimensión Producto contará con una jerarquía de Producto – Familia. Esto nos permitirá filtrar según familia de producto o por un producto en concreto.

La dimensión Centro Comercial contará con una jerarquía de “dos caminos”. De modo que podremos analizar por cada atributo de la jerarquía.
Estos atributos son los siguientes:
- Zona-Ciudad-Centro comercial
- Tipo - Centro comercial

Como se puede observar el campo ImpactPoint no ha sido añadido en esta jerarquía. En su lugar, se han creado los indicadores Impactos_BigScreen e Impactos_Information_Point_Screen descritos anteriormente.

Como se puede observar todas las tablas, excepto la tabla de hechos, contiene un campo VARCHAR donde introduciremos lo datos del nombre del atributo en concreto.

4.1.2. Cubo Visitas.

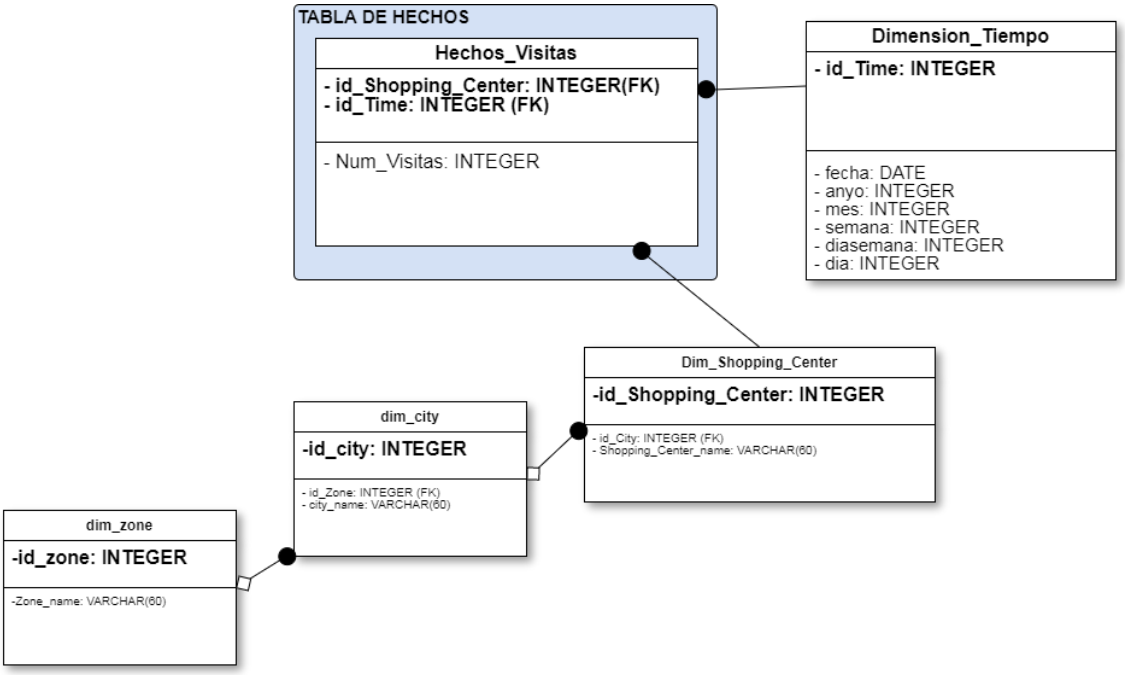


Figura 9. Detalle Diseño Cubo Visitas

Como podemos observar el cubo de Visitas es más sencillo. Para este caso nos basamos en dos dimensiones Tiempo y Centro Comercial.
Su tabla de hechos está compuesta por las Claves Primarias IdShoppingCenter e IdTiempo. Ambas claves ajenas hacen referencias a las claves primeras de las dimensiones ShoppingCenter y Tiempo.

Como en el caso del cubo de Ventas – Impactos, la dimensión Tiempo contará con una tabla cuya clave primaria será su idTiempo y con unos indicadores de fecha, año, trimestre, mes, semana y diasemana dado que se requiere una granularidad mínima.

En el caso de la Dimensión ShoppingCenter únicamente necesitaremos conocer la Ciudad y la Zona por lo que se ha creado una jerarquía de dos niveles para ello.

4.2. Implementación del DWH.

Tal y como anteriormente se ha comentado el DataWareHouse será un schema de MySQL. Este esquema recibe el nombre de datawarehouse. La implementación del DATAWAREHOUSE se reduce a la creación de las tablas correspondientes a las tablas de hechos y dimensiones definidas en la fase de diseño. A continuación, se muestra una imagen con la implementación de las tablas del schema DATAWAREHOUSE.

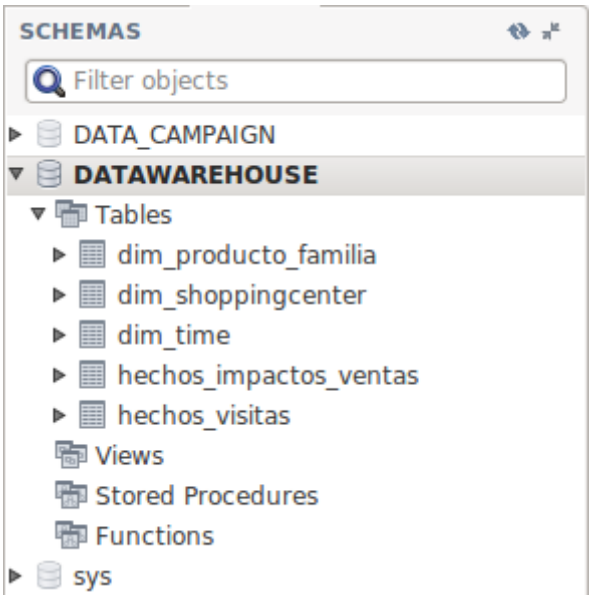
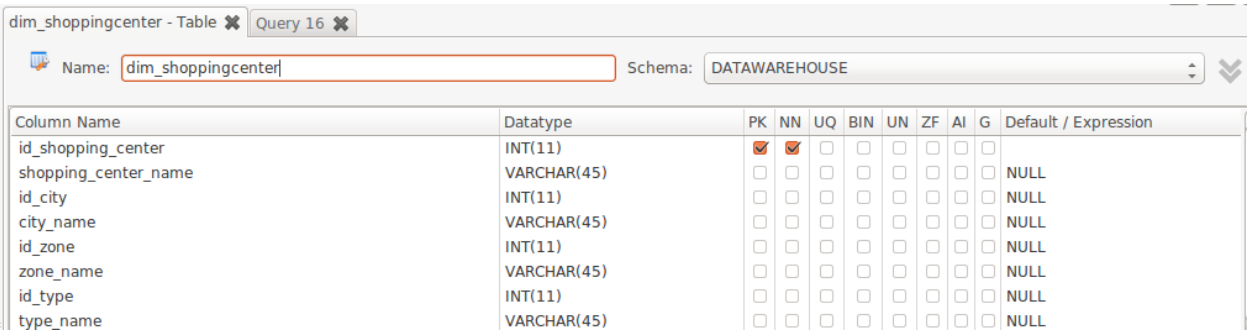


Figura 10. Tablas Datawarehouse

Como podemos observar en la imagen se han creado las dimensiones producto_familia, shopping_center y time. Por otro lado, tenemos las tablas de hechos de impactos-ventas y visitas.
En lo que respecta al detalle de cada una de las tablas a continuación se muestra las columnas y tipo de datos utilizados en cada caso, además de su correspondencia con los diseños presentados en el punto anterior.



Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	G	Default / Expression
id_shopping_center	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
shopping_center_name	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
id_city	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
city_name	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
id_zone	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
zone_name	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
id_type	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
type_name	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL

Figura 11. Tablas dim_shoppingcenter

Query 18 | dim_time - Table

Name: dim_time Schema: DATAWAREHOUSE

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	G	Default / Expression
id_time	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
date	DATE	☐	☐	☐	☐	☐	☐	☐	☐	NULL
year	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
month	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
weekofyear	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
weekday	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
day	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
dateiso	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL

Figura 12. Tablas dim_time

dim_shoppingcenter - Table | Query 16 | dim_producto_familia - Table

Name: dim_producto_familia Schema: DATAWAREHOUSE

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	G	Default / Expression
id_producto	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
nombre_producto	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
id_familia	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
nombre_familia	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL

Figura 13. Tablas dim_producto_familia

Query 18 | hechos_impactos_ventas - Table

Name: hechos_impactos_ventas Schema: DATAWAREHOUSE

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	G	Default / Expression
id_shopping_center	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
id_producto	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
id_tiempo	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
impactos_big_screen	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
impactos_information_point_screen	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL
ventas	VARCHAR(45)	☐	☐	☐	☐	☐	☐	☐	☐	NULL

Figura 14. Tabla hechos_impactos_ventas

Query 18 | hechos_visitas - Table

Name: hechos_visitas Schema: DATAWAREHOUSE

Column Name	Datatype	PK	NN	UQ	BIN	UN	ZF	AI	G	Default / Expression
id_time	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
id_shopping_center	INT(11)	☑	☑	☐	☐	☐	☐	☐	☐	
num_visitas	INT(11)	☐	☐	☐	☐	☐	☐	☐	☐	NULL

Figura 15. Tabla hechos_visitas

4.3. Fuentes de datos.

Antes de proceder con el diseño debemos tener en cuenta la fuente de los datos. En el caso de este proyecto para simular un entorno real se han cargado previamente los datos en un schema de BBDD llamado DATACAMPAIGN. No obstante, y con el fin de aprender el manejo de la herramienta se ha optado en algunos casos a cargar los datos directamente desde un fichero CSV generado a partir de las hojas de cálculo proporcionadas. Estos archivos CSV hacían referencia a las hojas concretas que a su vez serán las tablas que disponemos en el schema DATACAMPAIGN.

A continuación, se muestra una imagen con las tablas generadas. Como anteriormente se ha comentado la intención de este schema es el simular el acceso a datos de un entorno real.

Figura 16. SCHEMA DATACAMPAIGN

La siguiente imagen muestra los archivos CSV generados a partir de los ficheros Excel proporcionados.

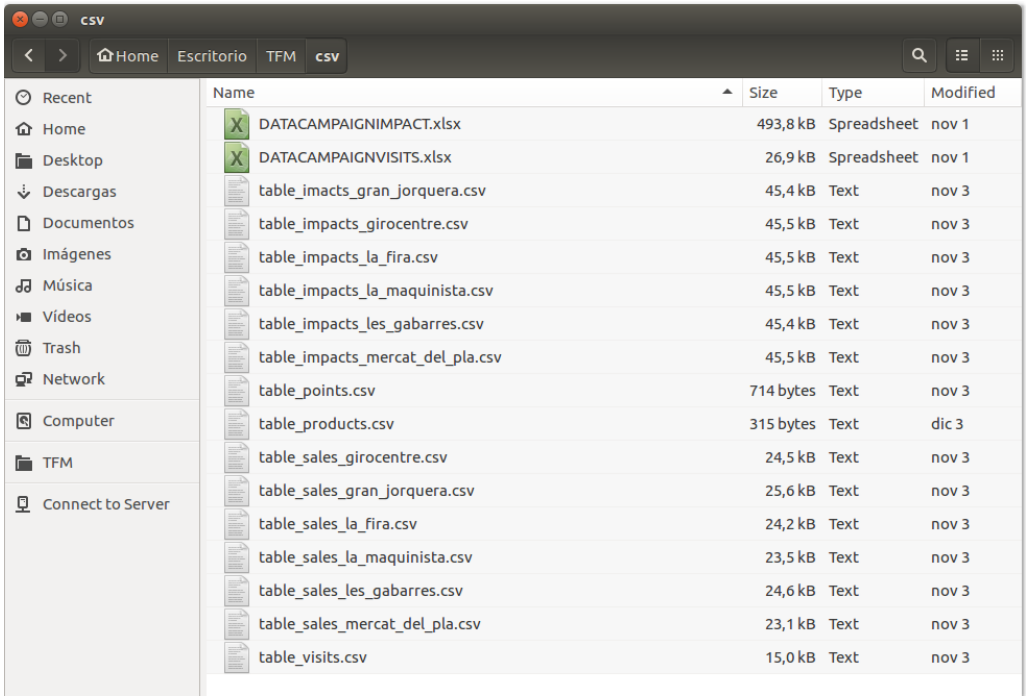


Figura 17. Ficheros de datos (CSV y Excel)

Tal y como anteriormente se ha comentado, en la implementación de los ETL se han utilizado ambas alternativas con la finalidad de comprobar su funcionamiento en cada caso.

5. Diseño e implementación de los procesos ETL

5.1. Introducción

Los sistemas o procesos ETL (Extact-Transform-Load) son la base de la construcción de cualquier sistema Data Warehouse (aunque además puedan ser utilizados para otros muchísimos cometidos). Un sistema bien diseñado extrae la información de los sistemas origen, asegura la calidad y consistencia de los datos, homogeniza los datos de sistemas divergentes para que puedan ser utilizados de una forma conjunta (procesando y transformando la información si es necesario) y finalmente genera los datos en el formato apropiado para que puedan ser utilizados por las herramientas de análisis.

El proceso de construcción de un sistema ETL puede ser extraordinariamente exigente y complejo, estando ademas limitado por muchos aspectos, como pueden ser los requerimientos, los formatos y deficiencias de los datos de origen, las habilidades del personal disponible, las necesidades de los usuarios finales, el presupuesto del proyecto, las ventanas de tiempo para los procesos de actualización, etc. Teniendo en cuenta esto, no se debe nunca despreciar la importancia, el tiempo y recursos que se han de utilizar para su construcción.

Existen diferentes elementos que van a afectar en cómo va ser o como vamos a construir nuestro sistema ETL. Los más importantes son los requerimientos. La elección de uno o varios procesos de negocio, las dimensiones e indicadores que vamos a analizar, su granularidad, etc., van a determinar cosas tan dispares como los orígenes de datos que vamos a tener que utilizar, la forma de procesar la información, la complejidad de los procesos, etc. Esto nos va a hacer darnos cuenta de lo importante que son dichos requerimientos y su correcta definición en todas las tareas que realicemos a continuación, incluyendo la definición de la arquitectura de nuestros procesos ETL. Además el estudio de los sistemas origen debe permitir añadir nuevas funcionalidades que pueden no haber sido tenidas en cuenta antes. [12]

Como explica muy bien Kimball en su libro, los procesos ETL son similares a un restaurante y su cocina. En el comedor, los comensales degustan los platos como lo harian los analistas de negocio con los datos utilizando sus correspondientes herramientas de análisis. Puertas atras, en el interior, en la cocina, se preparan los platos, se analizan y limpian los ingredientes, desechando aquellos que no estan en condiciones, se trocean, se cocinan, hasta elaborarlos tal y como serán presentados a los clientes.

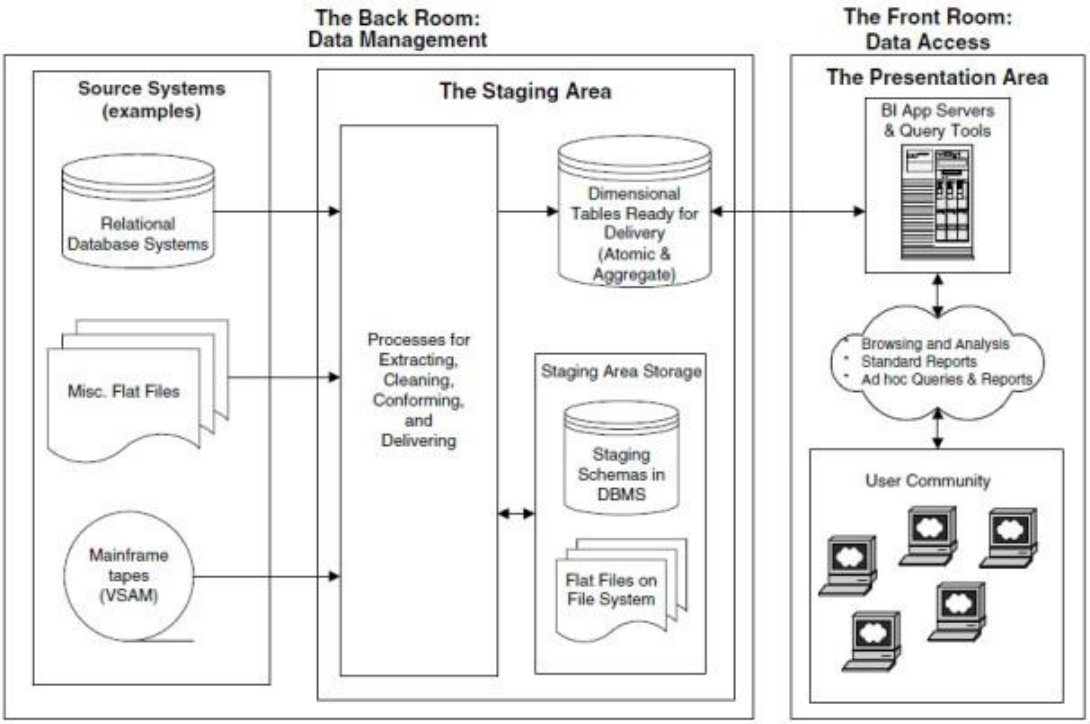


Figura 18. Area de Staging según kimball

De forma similar, el area de Stage sera lo mismo para nuestro DW. Es un lugar al que solo acceden las personas especializadas en la integración de datos, fuera del alcance de los usuarios. Allí los datos son extraidos, depurados, limpiados, conformados y normalizados, manipulados o calculados, y preparados para ser cargados en el DW donde podrán ser accedidos por los usuarios para realizar análisis con las diferentes herramientas de las que dispongan.

Los procesos en el area de Staging pueden incluir o no un almacenamiento de datos (aunque sea temporal), cuestión que dependera de cómo se diseñen los procesos, de los volumenes de información o de otras cuestiones. Básicamente, tenemos 4 tipos de pasos en esta area:

Extracción: los datos son extraidos de los sistemas origen, que pueden ser tanto bases de datos, como ficheros (estructurados o no) u otros origenes. Los procesos de extracción pueden ser a veces el lugar idóneo para realizar las primeras acciones sobre los datos, como formateo, conversiones de tipos, conversión de juegos de caracteres, etc.
Depuración: en esta etapa de los procesos ETL se procesa la calidad de los datos, revisando valores válidos, consistencia, eliminación de valores redundantes, chequeo de reglas complejas, etc. Puede ser necesaria la intervención humana en determinados casos.
Conformación/normalización: la información es unificada, conformada y normalizada. Los indicadores y ratios son calculados de una forma racional, lo mismo que los atributos de las dimensiones, para que esten unificados y en todos los sitios donde aparezcan tengan la misma estructura y el mismo significado.

Entrega: la información está- preparada p5ara ser analizada. Se entrega al DW para que las herramientas de análisis puedan utilizarla, en los formatos idoneos para dicha tarea.

5.2. Diseño ETL.

Para el diseño del ETL se ha utilizado la herramienta Pentaho Data Integration. Esta herramienta nos facilita de una forma simple e intuitiva la extracción, transformación y carga de los datos.

En el punto anterior, se ha explicado el proceso de diseño e implementación de los procesos ETL. En el caso de este proyecto, se ha tenido en cuenta, por un lado, la fuente de los datos a utilizar la cuál en unos casos se ha accedido a tablas de base de datos y en otros casos se ha accedido a los archivos csv. Por otro lado, se ha tenido en cuenta el diseño de los cubos definidos a partir de los cuales se han creado las tablas de dimensiones y hechos del datawarehouse. Por ejemplo, en el caso de la tabla de dimensión tiempo y la tabla de dimensión shopping center son comunes a los cubos diseñados, por lo que solo necesitaremos crear un ETL en cada caso para crear dichas dimensiones. En el caso, de la dimensión producto únicamente será válida para el cubo ventas impactos.

En lo que respecta a las tablas de hecho, se deberá tener en cuenta los identificadores “id” que en cada dimensión se han definido y crearemos un proceso ETL que se encargue de su creación. A continuación, se explica cada caso particular.

Para el caso de la tabla de hechos visitas dada su simplicidad será necesaria la creación de un proceso ETL. Esto se debe a que únicamente tenemos una fuente de datos, en este caso la tabla visitas de DATACAMPAIGN.

Para el caso de la tabla de hechos impactos – ventas, necesitaremos acceder a distintas fuentes de datos. Esto hace un poco más complicado su carga en el datawarehouse ya que la intención es que a partir de varias tablas o archivos csv obtengamos una tabla de hechos. En este caso, y para simplificar su creación se ha un ETL por cada shopping_center que se encargará de alimentar la tabla de hechos. Por otro lado, un Job se encargará de orquestar dicha carga de datos estableciendo un orden. Este orden de ejecución se ha establecido para que en caso de tener necesidad de cargar nuevos datos cada vez, sea únicamente un ETL el que se encargue del vaciado de la tabla antes de su carga.

En el siguiente punto se muestran los distintos ETLs creados. En el Anexo 2 de este documento podemos ver la relación de ETLs creados para este proyecto. Ademas, podemos ampliar el detalle de los distintos archivos ETL revisando los anexos del 2 al 11 de este documento.

5.3. Implementación del ETL.

5.3.1. Hechos_Impactos_Ventas

Comenzaremos explicando cómo se ha creado la tabla de hechos impactos – ventas. Para la creación de esta tabla es necesario el acceso a las tablas de datos con los impactos realizados en cada centro comercial y por otro lado las ventas realizadas en cada centro comercial. La intención final es la de disponer una única tabla de hechos que nos permita explotar los datos para su análisis en el BI. Dada su complejidad se ha decidido dividir la tarea en 6 procesos ETL, uno por cada centro comercial e implementar un job que se encargue de orquestar la carga de datos de los distintos ETL generados.

La figura 23 muestra cómo se cargan los datos de Ventas e impactos del centro comercial la Gran Jorquera. Dado que la tabla impactos_gran_jorquera dispone información de los impactos en Big Screen y en Information Point Screen se ha creado una consulta a la misma tabla por cada uno de los criterios. Por otro lado, se ha realizado una consulta a la tabla ventas_gran_jorquera. En el proceso se añaden los ids necesarios para su enlace con el resto de dimensiones y finalmente se cargan en la tabla del datawarehouse hechos_impactos_ventas.

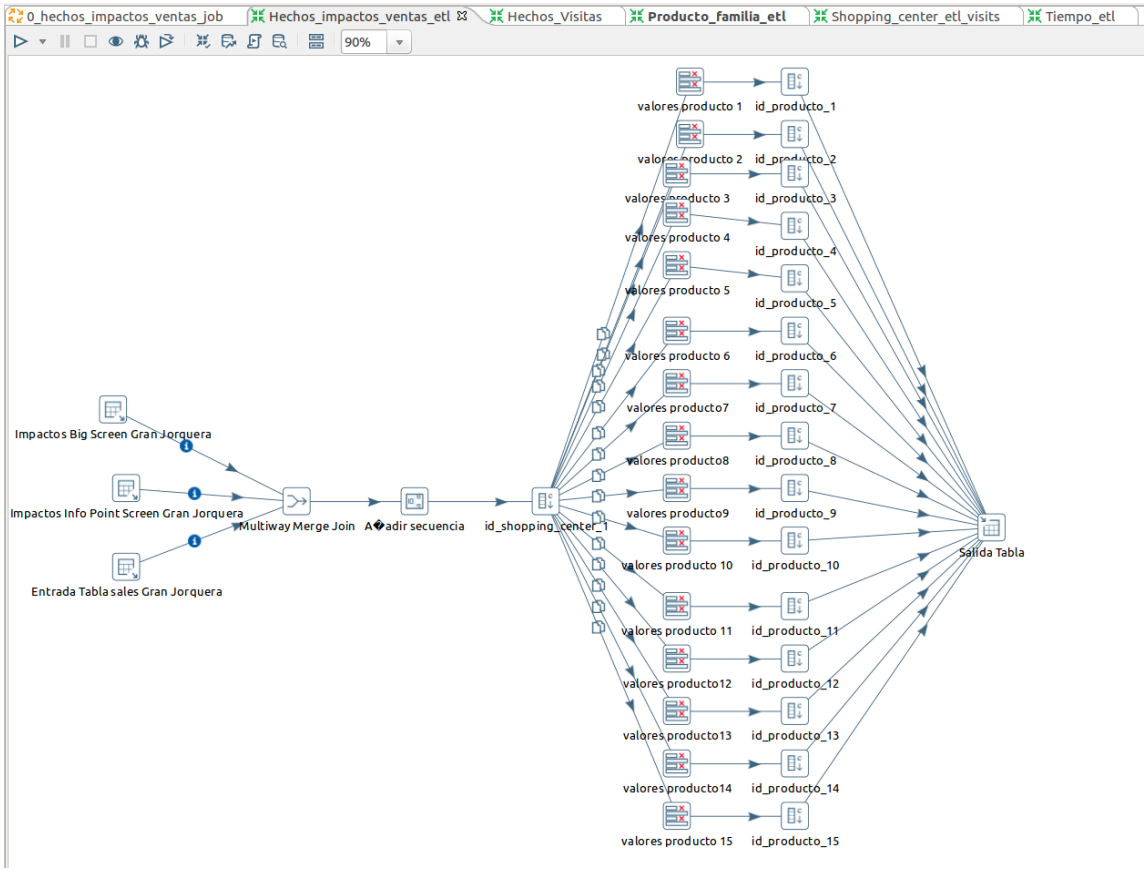


Figura 19. Ejemplo ETL hechos_impactos_ventas_Gran_Jorquera

Tal y como anteriormente se ha comentado, para el caso de la tabla de hechos impactos – ventas, necesitaremos acceder a distintas fuentes de datos. Esto hace un poco más complicado su carga en el datawarehouse ya que la

intención es que a partir de varias tablas o archivos csv obtengamos una tabla de hechos. En este caso, y para simplificar su creación se ha un ETL por cada shopping_center que se encargará de alimentar la tabla de hechos. Por otro lado, un Job se encargará de orquestar dicha carga de datos estableciendo un orden. Este orden de ejecución se ha establecido para que en caso de tener necesidad de cargar nuevos datos cada vez, sea únicamente un ETL el que se encargue del vaciado de la tabla antes de su carga. La figura 24 muestra el proceso.

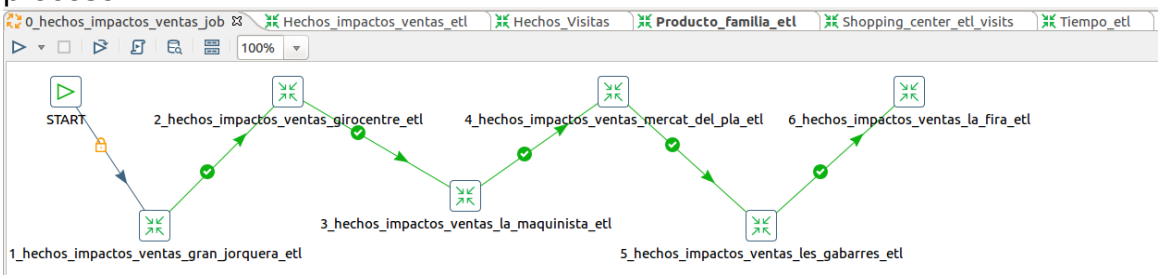


Figura 20. Job tabla de hechos impactos-ventas

5.3.2. Hechos_Visitas

La tabla hechos_visitas tal y como se ha visto en el punto anterior, en el datawarehouse, está compuesta por los valores id_shopping_center, id_time y visitas. En este caso, simulando un entorno real se ha procedido a acceder a los datos de la tabla Visits de la base de datos DATA_CAMPAIGN. Una vez obtenidos los datos se ha cargado una secuencia con el número de día del año en cuestión el cuál hace referencia al id_time, con el que enlazaremos en el cubo con la dimensión tiempo. El siguiente paso ha sido añadir el id correspondiente al centro comercial junto con las visitas de ese centro comercial mediante la función selecciona/renombra valores y añadir constante. Por último, se cargan los datos en la tabla de hechos_visitas del datawarehouse. A continuación, se muestra la imagen con el workflow del ETL:

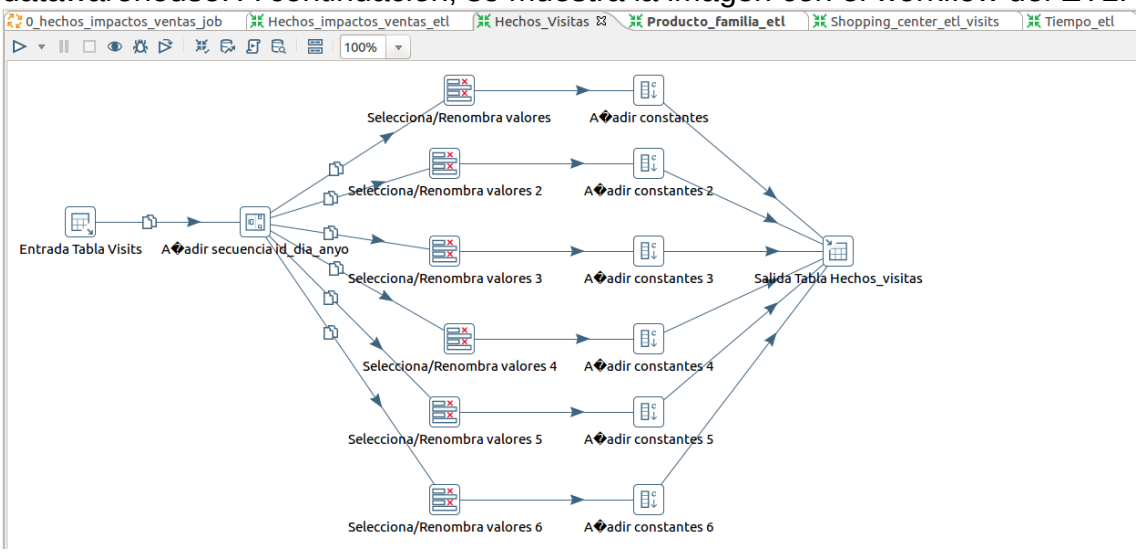


Figura 21. ETL tabla de hechos_visitas

5.3.3. Dimensión Producto y Familia

Para la creación de la dimensión producto y familia se ha procedido a la carga de datos desde el csv productos el cual estaba compuesto por el campo producto y el campo familia en cuestión, ambos tipo string. Para su transformación mediante la función filas únicas se ha filtrado los distintos campos familia existentes y se le ha añadido un id, en paralelo dado que el csv contenía todos los productos se ha añadido un id que haga referencia al mismo. Una vez obtenidos los id_familia e id_producto se ha realizado un Merge Join para unir los valores y cargarlos finalmente en el datawarehouse en la tabla dim_producto_familia. A continuación se muestra la imagen con el ETL:

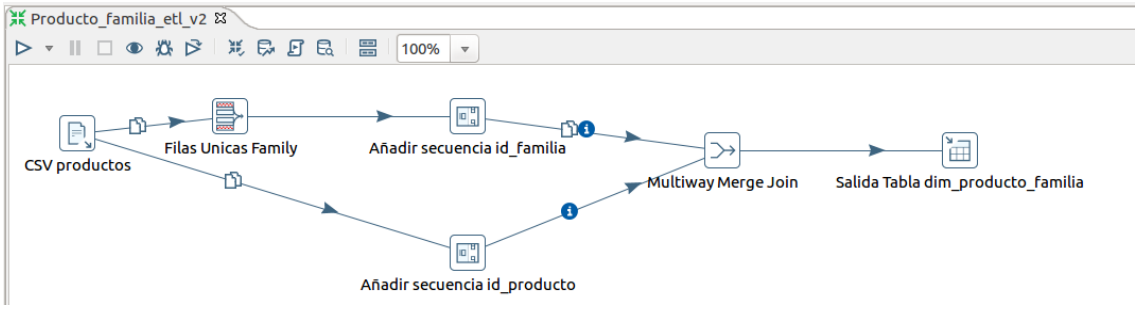


Figura 22. ETL dimensión producto y familia

5.3.4. Dimensión Shopping_Center

La dimensión shopping_center ha sido junto con la tabla de hechos impactos_ventas la más compleja a la hora de transformar los datos. En este caso partimos del csv Points compuesto por los string zone,city, shopping center, type e impact point. Lo que se pretende en esta dimensión es el obtener una referencia o id para cada uno de los valores, a su vez, el id_shopping_center será utilizado como clave ajena en la tabla de hechos impactos-ventas y hechos_visitas.

Para ello, se han cargado los datos a partir del csv en cuestión (tabla_points.csv) y se ha procedido a su mapeo, diviendo en 4 flujos paralelos su tratamiento para generar cada uno de los ids. A medida que se han ido añadiendo los ids podemos observar como mediante los Merge Join se han ido uniendo los resultados para finalmente cargarlos en la tabla de hechos del datawarehouse.

A continuación, se muestra la imagen con la carga de datos de esta dimensión.

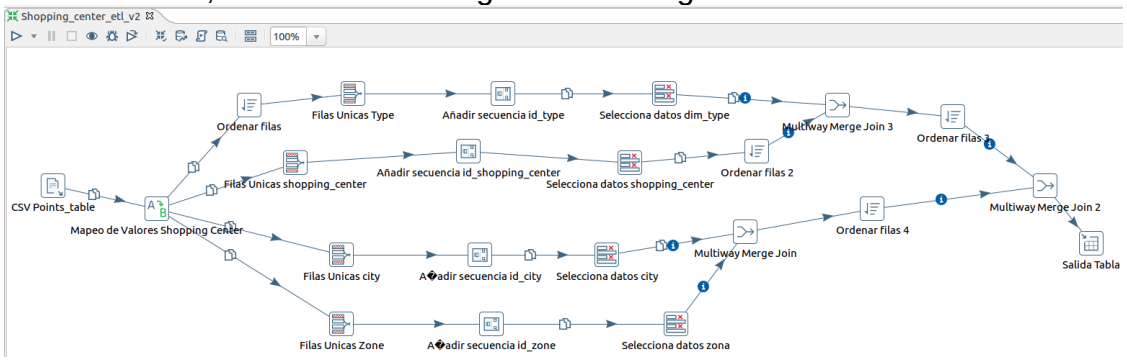


Figura 23. ETL dimensión jerarquica shopping_center

5.3.5. Dimensión Tiempo

Para la dimensión tiempo se ha optado por crear un flujo que genere y calcule los días del año. Para ello en primer lugar se van generando las filas a las cuales se les añade una secuencia de números, a partir de estos números se calcula la fecha mediante la función Calculadora del spoon. El siguiente paso elimina los campos de ayuda y se procede a calcular la información de los días. Limitamos al año 2017 la salida a la tabla dimesión dim_time. A continuación se muestra el flujo.

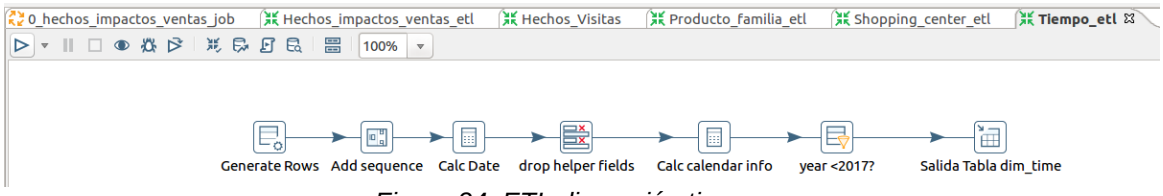


Figura 24. ETL dimensión tiempo

6. Creación del Entorno BI para el análisis

6.1. Creación de los cubos

Para la creación de los cubos se ha utilizado la herramienta de Pentaho Schema Workbench.

Antes de comenzar con su creación se ha revisado el diseño definido en el punto 4 de este documento identificando si tenemos dimensiones comunes en ambos cubos (visitas y ventas-impactos). Schema Workbench permite crear dimensiones a nivel de cubo o a nivel de schema, en el caso de crear una dimensión a nivel de cubo esa dimensión estará únicamente definida en ese cubo, en el caso de crear la dimensión a nivel de schema nos permitirá ser usada en distintos cubos lo que nos ahorrará el tener que hacer trabajo por duplicado. Para este proyecto, se identifican dos dimensiones comunes en ambos cubos, la dimensión tiempo y la dimensión shopping_center. Por otro lado, la dimensión producto-familia es utilizada únicamente en el cubo Ventas-impactos aunque se ha dado continuidad al criterio de crear a nivel de schema.

La imagen 25 muestra el diseño del cubo con las jerarquías definidas en cada una de las dimensiones.

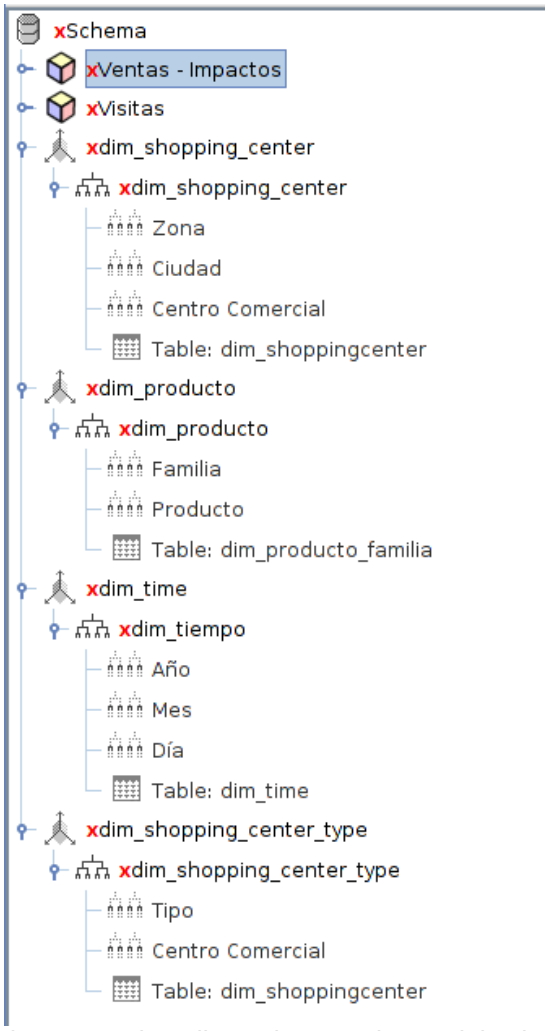


Figura 25. Vista dimensiones Schema del cubo.

Como podemos observar en la definición de las jerarquías de cada una de las dimensiones se ha creado de un nivel global a un nivel de más detalle en cada uno de los casos. Por ejemplo, en el caso de la dimensión tiempo hemos comenzado por Año, siguiente nivel será el mes y por último el día. En el caso del shopping_center se han creado dos dimensiones una para el tratamiento por tipo y otra para el tratamiento a partir de zona tal y como diseñamos en los cubos.

Una vez definidas las dimensiones se procede a crear los cubos. Para la utilización de las dimensiones anteriormente descritas se ha utilizado la opción dentro del cubo de “Add Dimension Usage” y seleccionando la dimensión que deseamos utilizar en cada caso. Una vez añadidas las dimensiones se ha procedido a añadir la tabla de hechos sobre la cual se basan los datos a analizar, seleccionando en cada caso la que pertenece al cubo correspondiente. Por último, se han añadido las métricas necesarias para el análisis de los datos. La siguiente imagen muestra el detalle de cada uno de los cubos definidos.

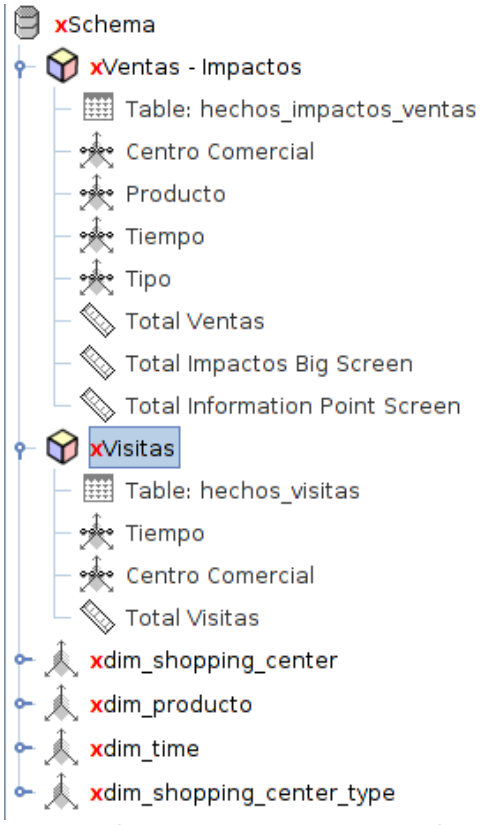


Figura 26. Cubo Ventas –Impactos y Cubo Visitas.

Como observación, vemos que los cubos y dimensiones aparecen con un aspa en rojo, esto es un mensaje de advertencia debido a que no está definido el schema utilizado en cada uno ellos, únicamente se hace referencia a las tablas en cuestión ya que la conexión se establece a la base de datos mysql DATAWAREHOUSE directamente y no es necesario indicar el schema de la base de datos en este caso.

Podemos ampliar el detalle de los cubos generados revisando el Anexo 12 de este documento.

6.2. Diseño e implementación de los paneles de mando.

En esta fase del proyecto se procede a la creación de los paneles de mando que permita de una manera sencilla responder a las cuestiones planteadas.
Es importante en esta fase tener claras las cuestiones y su respuesta ya que de su resultado depende el éxito del proyecto.

Para el diseño e implementación de los paneles de mando se ha optado por la creación de queries independientes mediante saiku las cuales luego utilizaremos en los paneles de mando.

De modo que para responder a cada una de las cuestiones planteadas se ha creado un query los cuales se explicarán con más detalles a continuación.

En primer lugar, tal y como se ha indicado anteriormente, debemos tener claras las cuestiones planteadas a las que debemos dar respuesta. En este caso son las siguientes:

- *En general, ¿la publicidad basada en impactos visuales, es efectiva?*
- *¿Hay determinadas zonas geográficas donde los impactos son más efectivos?*
- *Según la familia de producto, ¿existe alguna relación entre los artículos impactados y el resultado obtenido?*
- *¿Son más efectivos los impactos según el momento temporal en el que se producen?*
- *¿Existe algún artículo o familia que según la zona, su publicidad tiene mayor impacto? ¿Y según el momento cuando se realiza?*

Para justificar las respuestas planteadas necesitamos apoyarnos en los datos obtenidos y adaptarlos según precisemos. Tras haber realizado los pasos previos de extracción, transformación y carga y la creación de los cubos procederemos a realizar el último paso del proceso, la creación de las consultas.
En el siguiente punto de este documento se realizará el análisis de las cuestiones planteadas dando respuesta a cada una de las cuestiones a partir de las distintas cuestiones y paneles de mando generados.

7. Análisis de la información

7.1. Análisis de los datos.

El objetivo de este proyecto, es el de dar respuesta a una serie de cuestiones planteadas a partir de una información, simulando un entorno real de una campaña publicitaria realizada en distintos centros comerciales. Estas cuestiones deben ser respondidas a partir de la información obtenida durante la campaña, a través de la herramienta Business Intelligence.

En este punto vamos a responder las cuestiones planteadas a través de las "queries" y paneles de mando creados. Para su explicación se propone en primer lugar la cuestión planteada y acto seguido se mostrá la query realizada junto con una explicación y justificación de la respuesta.

Para todas las cuestiones planteadas se ha utilizado el cubo Impactos-Ventas ya que es el que contiene la información para dar respuesta a dichas cuestiones.

A continuación, en primer lugar responderemos a las cuestiones a través del cubo Impactos – Ventas. Acto seguido, se mostrarán algunas queries realizadas a través del cubo Visitas y finalmente se mostrarán algunos paneles de mando generados.

7.2. Queries Impactos-Ventas.

- ***En general, ¿la publicidad basada en impactos visuales, es efectiva?***

  Al tratarse de una cuestión que intenta calificar de una forma general si es efectiva o no la publicidad basada en impactos antes de sacar conclusiones vamos a en primer lugar analizar bien los datos respondiendo a las demás preguntas planteadas antes de dar una respuesta.
  Esto nos permitirá dar una respuesta más acertada con respecto a la efectividad de los impactos con relación a las ventas realizadas. En el punto 7.2 de este documento dedicado a las conclusiones se responderá a esta cuestión.

- ***¿Hay determinadas zonas geográficas donde los impactos son más efectivos?***

  La siguiente consulta (figura 28) se plantea a nivel de zona geográfica. Si observamos los resultados vemos como a simple vista hay un menor número de impactos en la Zona Centro, sin embargo, sus ventas han sido mayores que en el Norte o en Sur, por lo que podemos concluir que SI haz zonas geográficas donde los impactos son más efectivos, para ser más concretos en la zona Centro.

Zona	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
CENTER	8.813.596	54.129	53.957
NORTH	7.207.765	56.435	56.376
SOUTH	5.442.174	56.218	56.172

Figura 27. Efectividad de los impactos por zona.

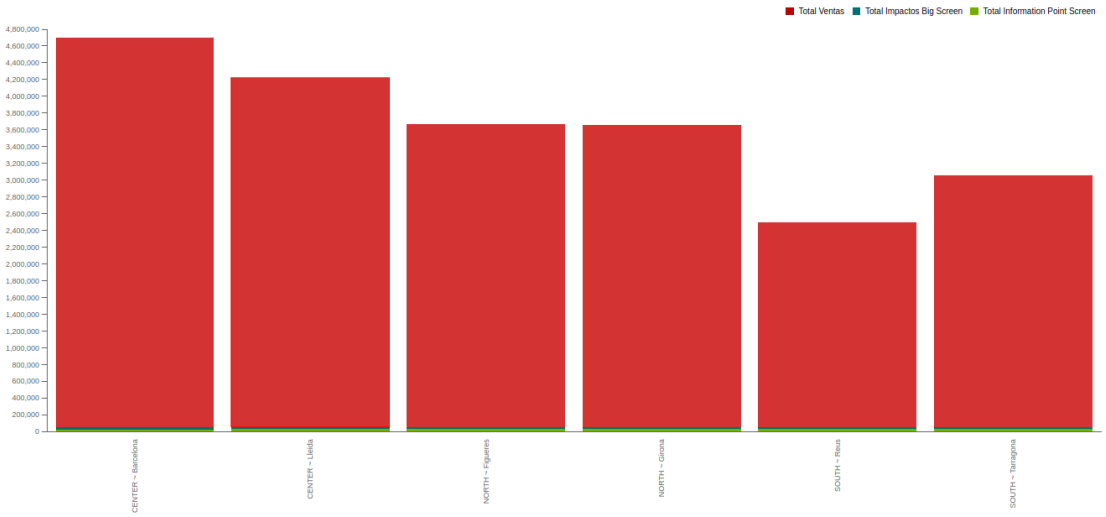
Como podemos observar si quisiésemos responder a la primera pregunta a partir de esta consulta no podríamos dar una respuesta clara ya que, como ya se ha mencionado, cada consulta debemos adaptarla a la cuestión que estamos planteando.

Para conocer con más detalle la efectividad de dichas ventas se han añadido las ciudades que pertenecen a cada una de las zonas para comprobar dicha efectividad. La siguiente imagen muestra cómo se confirma cómo en la ciudad de Barcelona se han efectuado más ventas con respecto con un menor número de impactos, en comparación con el resto de ciudades. Este mayor número de ventas podríamos atribuirlo también a que Barcelona es una ciudad más poblada con respecto al resto, y que no todas las ventas estén relacionadas con el número de impactos. Por lo que a partir de esta segunda consulta, no podríamos decir con certeza que hay una efectividad existente entre el número de ventas y los impactos según la zona geográfica.

Zona	Ciudad	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
CENTER	Barcelona	4.648.187	25.862	25.862
	Lleida	4.165.409	28.267	28.095
NORTH	Figueres	3.609.643	28.227	28.178
	Girona	3.598.122	28.208	28.198
SOUTH	Reus	2.442.996	28.117	28.137
	Tarragona	2.999.178	28.101	28.035

Figura 28. Efectividad de los impactos por ciudad

La siguiente imagen (figura 30) pretende mostrar los datos en modo gráfico para su comparación pero debido a la gran diferencia entre las cifras de ventas y los impactos no podemos ver con claridad esa influencia de los impactos con respecto a las ventas según la ciudad.



. *Figura 30. Gráfico Efectividad de los impactos Consulta3.*

- ***Según la familia de producto, ¿existe alguna relación entre los artículos impactados y el resultado obtenido?***

Para responder a esta cuestión debemos realizar una consulta que nos permita visualizar el total de impactos y ventas agrupados por familias de productos.

Como podemos observar en las figuras 30 y 31 destaca la venta de aparatos electrónicos. En concreto, las televisiones tienen un gran número total de ventas con respecto al resto de productos. Si observamos el total de impacto a nivel de familias de productos no vemos significantes diferencias ya que en su suma es más o menos igual en todos los casos. Si entramos más en el detalle de los productos, podemos observar como hay productos los cuales tienen un notable número de impactos pero sus ventas no han sido excesivamente buenas y viceversa.

Familia	Producto	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
Food	PIZZA TDRL	214.101	9.029	9.071
	PIZZA DROE	247.546	10.893	11.003
	PIZZA CUSN	291.084	13.717	13.746
Electronics	TV SMRT42SNG	4.549.278	9.036	8.953
	TV SMRT42PHI	5.132.042	11.011	10.921
	TV SMRT42SNY	5.925.139	13.773	13.735
Clothing	TROUSERS LVS	344.258	9.113	9.038
	TROUSERS MNG	386.938	10.980	11.056
	TROUSERS LEE	441.843	13.715	13.687
Sports	SHOES NKE	891.406	9.116	9.037
	SHOES ADS	1.038.346	10.954	11.041
	SHOES ACS	1.253.045	13.786	13.677
Home	FRAME 18X10 CK	240.168	9.092	9.076
	FRAME 18X10 CH	253.657	11.061	10.993
	FRAME 18X10 MD	254.684	11.506	11.471

Figura 29. Relación familias de productos, artículos impactados y ventas

Si comparamos por ejemplo el producto PIZZA CUSN, el producto TROUSERS LEE y la TV SMRT42SNY, vemos como los impactos realizados en estos productos son similares sin embargo hay una gran diferencia entre PIZZA CUSN y TROUSERS LEE, la cual se dispara cuando observamos el producto TVSMRT425SNY.

Familia	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
Food	752.731	33.639	33.820
Electronics	15.606.459	33.820	33.609
Clothing	1.173.039	33.808	33.781
Sports	3.182.797	33.856	33.755
Home	748.509	31.659	31.540

Figura 30. Relación familias impactos y ventas

Por otro lado, como anteriormente se ha comentado si comparamos el número de impactos de los aparatos Electrónicos y sus ventas vemos una gran diferencia de ventas con respecto al resto de familias de productos.

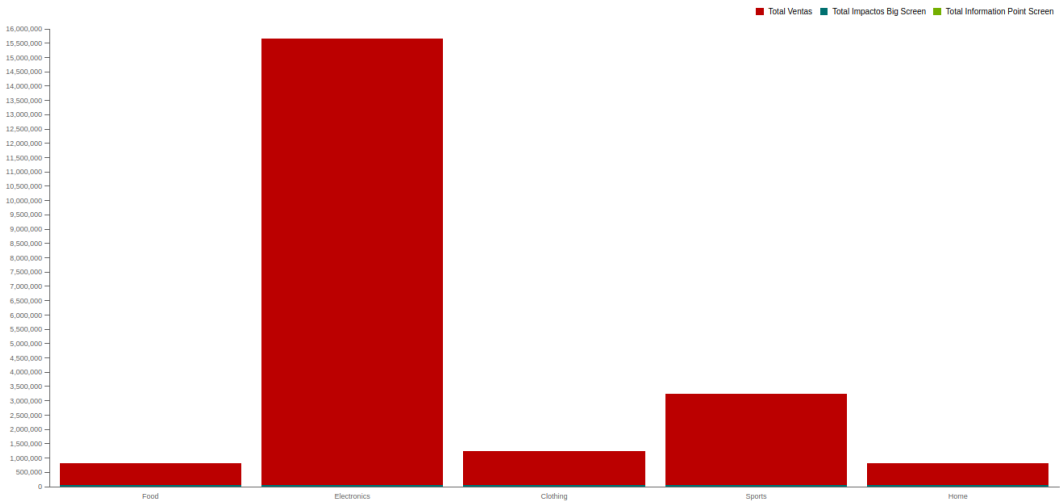


Figura 31. Relación familias impactos y ventas

El gráfico anterior nos muestra claramente esta diferencia comentada entre las ventas de cada una de las familias. Por lo que podríamos concluir, que la campaña de impactos ha sido muy positiva para la venta de aparatos electrónicos, en este caso, la venta de Televisiones de varias marcas ha sido muy buena alcanzado la cifra de más de 15millones de ventas. En segunda posición se situan los productos deportivos, para estos también podríamos decir que la campaña de impactos ha influido positivamente en las ventas. En el caso del resto de productos, no ha sido tan efectiva ya que con un número similar de impactos tenemos un número de ventas bastante inferior que los mencionados.

- ***¿Son más efectivos los impactos según el momento temporal en el que se producen?***

Para responder a esta cuestión se ha planteado una sencilla consulta en la que veamos el total de impactos y ventas para comparar los resultados por tiempo y comprobar su efectividad.

Año	Mes	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
2016	1	2.051.328	16.939	16.892
	2	1.922.587	15.780	15.797
	3	2.074.531	16.848	16.891
	4	1.413.382	5.386	16.579
	5	1.478.987	5.493	17.258
	6	1.449.243	5.311	16.524
	7	2.035.377	16.909	16.927
	8	2.063.509	16.895	16.841
	9	2.027.513	16.308	16.414
	10	1.637.560	17.129	5.554
	11	1.602.730	16.611	5.278
	12	1.706.788	17.173	5.550

Figura 32. Efectividad de los impactos por tiempo.

La imagen anteriormente mostrada (figura 27) se plantea una visión del total de ventas e impactos realizados cada mes. A partir de esta primera consulta podemos comprobar como hay una gran diferencia en cifras entre el número total de ventas y el número total de Impactos. No obstante, podemos apreciar como las ventas se resienten ligeramente cuando los impactos se ven reducidos. Podemos observar como los meses con mayor número de ventas, alrededor de los 2,000.000 (Enero, Febrero, Marzo, Julio, Agosto y Septiembre) son los que mayor número de impactos poseen (más de 30000 si sumamos los impactos de Big Screen y los impactos de los Point Screen). Mientras que los meses de Abril, Mayo y Junio vemos como el Total Impactos en Big Screen se ha reducido mientras que los impactos en los Information Point se han mantenido, esto ha provocado un total de 21000 a 22000 Impactos, con unas ventas de menos de 1,500.000, esta reducción también la comprobamos en los meses de octubre, noviembre y diciembre. De modo que, a partir de esta primera consulta podemos concluir que la publicidad basada en impactos tiene un grado de efectividad respecto a las ventas. Otra cuestión, sería entrar a valorar si el número de ventas justificaría la campaña publicitaria.

La siguiente gráfica confirma lo comentado anteriormente.

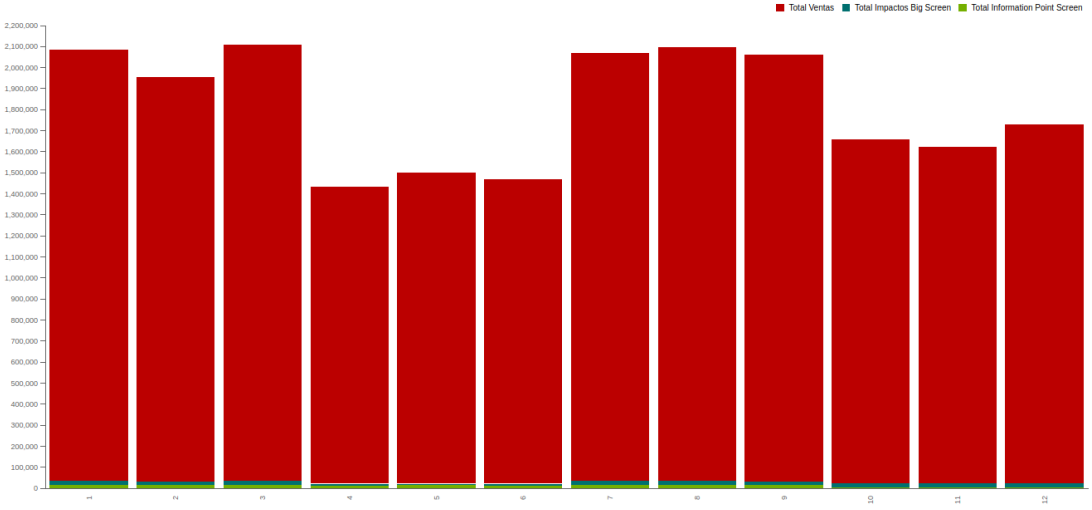


Figura 33. Gráfico Efectividad de los impactos por tiempo.

- ***¿Existe algún artículo o familia que según la zona, su publicidad tiene mayor impacto? ¿Y según el momento cuando se realiza?***

Para responder a esta cuestión se han planteado varias consultas a través del BI. Esta primera consulta (figura 35) muestra de una manera sencilla la relación de familia de productos, zonas y el total de impactos y ventas realizados.
A simple vista podemos sacar las siguientes conclusiones con respecto a la relación de la publicidad y su impacto en las ventas:

- Los productos de la familia *Food* con un número de impactos similar en todas las zonas, tiene un mayor número de ventas en la zona *Sur*.

- Los productos de la familia *Electronics* con un número de impactos similar en todas las zonas, tiene unas ventas superiores en la zona *Centro*.
- Los productos de la familia *Clothing* con un número similar de impactos en todas las zonas, tiene un mayor número de ventas en la zona *Centro*.
- Los productos de la familia *Sports* con un número similar de impactos en todas las zonas, tiene un mayor número de ventas en la zona *Sur*.
- Los productos de la familia *Home*, tienen un ligero número de impactos inferior en la zona Centro aun así sus ventas son ligeramente inferiores a la zona Norte y Sur con un mayor número de impactos. Entre la zona Norte y la zona Sur con un número similar de impactos ambas, tendría un mayor número de ventas la zona *Norte*.

	CENTER			NORTH			SOUTH		
Familia	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
Food	102.381	11.212	11.292	252.290	11.253	11.243	398.060	11.174	11.285
Electronics	7.763.910	11.315	11.195	5.098.344	11.309	11.295	2.744.205	11.196	11.119
Clothing	528.134	11.253	11.314	478.455	11.271	11.228	166.450	11.284	11.239
Sports	199.674	11.304	11.196	1.103.365	11.294	11.303	1.879.758	11.258	11.256
Home	219.497	9.045	8.960	275.311	11.308	11.307	253.701	11.306	11.273

Figura 34. Efectividad zona familia.

La siguiente figura mostraría la relación añadiendo el desglose por productos. Donde se confirman los datos mostrados anteriormente. No obstante, en esta vista vemos con más claridad como si hay una diferencia significativa con respecto a la efectividad de los impactos en la familia de productos Home en la zona Centro, donde se confirma su baja eficacia.

		CENTER			NORTH			SOUTH		
Familia	Producto	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
Food	PIZZA TDRL	33.987	3.007	3.041	72.269	3.005	3.026	107.845	3.017	3.004
	PIZZA DROE	34.165	3.600	3.664	82.105	3.694	3.648	131.276	3.599	3.691
	PIZZA CUSN	34.229	4.605	4.587	97.916	4.554	4.569	158.939	4.558	4.590
Electronics	TV SMRT42SNG	2.121.120	3.002	2.970	1.521.169	3.046	3.035	906.989	2.988	2.948
	TV SMRT42PHI	2.533.950	3.682	3.636	1.682.160	3.698	3.702	915.932	3.631	3.583
	TV SMRT42SNY	3.108.840	4.631	4.589	1.895.015	4.565	4.558	921.284	4.577	4.588
Clothing	TROUSERS LVS	152.406	3.037	3.016	136.790	3.044	3.040	55.062	3.032	2.982
	TROUSERS MNG	174.151	3.672	3.738	157.385	3.644	3.643	55.402	3.664	3.675
	TROUSERS LEE	201.577	4.544	4.560	184.280	4.583	4.545	55.986	4.588	4.582
Sports	SHOES NKE	66.114	3.053	2.992	308.962	3.017	3.022	516.330	3.046	3.023
	SHOES ADS	67.342	3.644	3.675	356.466	3.708	3.689	614.538	3.602	3.677
	SHOES ACS	66.218	4.607	4.529	437.937	4.569	4.592	748.890	4.610	4.556
Home	FRAME 18X10 CK	68.189	3.078	3.005	88.717	3.007	3.033	83.262	3.007	3.038
	FRAME 18X10 CH	78.152	3.673	3.664	91.477	3.669	3.675	84.028	3.719	3.654
	FRAME 18X10 MD	73.156	2.294	2.291	95.117	4.632	4.599	86.411	4.580	4.581

Figura 35. Efectividad zona familia y producto.

Con respecto a la eficacia publicitaria con respecto a la fecha y las zonas podemos sacar las siguientes conclusiones.

- La una relación de impactos similar en todas las zonas en las mismas fechas.
- Las ventas siempre son superiores en la Zona Centro.
- Si analizamos cada zona de manera independiente vemos una mayor eficacia el primer trimestre del año y los meses de verano (Julio, Agosto y Septiembre) dado que dichos meses tienen un mayor número de impactos.

		CENTER			NORTH			SOUTH		
Año	Mes	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen	Total Ventas	Total Impactos Big Screen	Total Information Point Screen
2016	1	879.157	5.440	5.422	682.292	5.827	5.719	489.879	5.672	5.751
	2	807.957	5.021	5.073	644.200	5.372	5.357	470.430	5.387	5.367
	3	887.200	5.429	5.430	691.347	5.710	5.736	495.984	5.709	5.725
	4	549.088	1.773	5.465	469.670	1.794	5.586	394.624	1.819	5.528
	5	572.937	1.827	5.742	495.959	1.859	5.738	410.091	1.807	5.778
	6	576.307	1.744	5.441	475.282	1.760	5.588	397.654	1.807	5.495
	7	868.203	5.445	5.398	673.135	5.766	5.782	494.039	5.698	5.747
	8	876.720	5.388	5.391	688.210	5.780	5.729	498.579	5.727	5.721
	9	857.558	5.236	5.252	673.108	5.552	5.606	496.847	5.520	5.556
	10	637.507	5.631	1.790	573.018	5.715	1.874	427.035	5.783	1.890
	11	611.290	5.503	1.753	563.414	5.574	1.789	428.026	5.534	1.736
	12	689.672	5.692	1.800	578.130	5.726	1.872	438.986	5.755	1.878

Figura 36. Efectividad zona y tiempo.

7.3. Queries Visitas.

Aunque no se solicita en el TFM es interesante de cara a una supuesta “toma de decisiones” tener en cuenta las visitas que ha habido en los centros comerciales y su relación con el número de impactos y ventas. A continuación se muestran algunas consultas donde podemos ver el número de visitas realizadas en los distintos centros comerciales de las distintas ciudades de los que disponemos datos.

		CENTER		NORTH		SOUTH	
		Barcelona	Lleida	Figueres	Girona	Reus	Tarragona
		La Maquinista	Mercat del Pla	Gran Jonquera Outlet & Shopping	Girocentre	La Fira	Les Gabarres
Año	Mes	Total Visitas	Total Visitas	Total Visitas	Total Visitas	Total Visitas	Total Visitas
2016	1	359.614	197.118	222.984	226.287	204.443	365.822
	2	350.353	180.842	185.416	208.213	212.405	348.541
	3	378.052	183.187	252.019	215.121	212.283	376.522
	4	343.774	185.185	213.020	209.703	207.563	359.877
	5	388.704	188.479	205.988	229.813	226.300	388.534
	6	361.293	198.463	223.905	219.675	204.070	361.271
	7	370.697	195.117	203.542	239.437	212.944	373.017
	8	374.440	198.562	224.188	208.541	211.175	383.210
	9	373.461	190.331	211.536	216.917	227.253	374.507
	10	360.084	201.851	214.293	204.554	219.986	373.322
	11	373.075	183.374	217.644	205.674	222.321	356.609
	12	373.319	188.882	214.896	225.713	206.249	369.450

Figura 37. Detalle Visitas.

Como podemos observar en la query, vemos como los centros comerciales con más visitas son La Maquinista de Barcelona y Les Gabarres de Tarragona. El menos visitado es el Mercat del Pla de Lleida
Observamos también como el número de visitantes se mantiene más o menos todo el año en cada uno de los centros comerciales.

La siguiente imagen muestra el total de visitantes por mes entre todos los centros comerciales.

Año	Mes	Total Visitas
2016	1	1.576.268
	2	1.485.770
	3	1.617.184
	4	1.519.122
	5	1.627.818
	6	1.568.677
	7	1.594.754
	8	1.600.116
	9	1.594.005
	10	1.574.090
	11	1.558.697
	12	1.578.509

Figura 38. Total Visitas por mes.

La imagen anterior confirma lo comentado anteriormente, el número de visitantes es mas o menos constante durante todo el año, en el cómputo global del mes.

No obstante, si observamos elt total de visitas día a día, si que podemos observar grandes diferencias en el número de visitantes entre unos días y otros. Como ejemplo, en la siguiente imagen podemos ver como el 1 de Enero o el 8 de Enero son días con más de 90000 visitantes, esto puede atribuirse al inicio de rebajas, sin embargo el día 7 de Enero no alcanzan ni las 50000 visitas, lo cuál se podría atribuir a que se trata del día de Reyes y normalmente las familias lo pasan en casa.

Año	Mes	Día	Total Visitas
2016	1	2016-01-01	91.208
		2016-01-02	0
		2016-01-03	57.207
		2016-01-04	56.371
		2016-01-05	55.108
		2016-01-06	52.823
		2016-01-07	46.158
		2016-01-08	93.806
		2016-01-09	0
		2016-01-10	56.525
		2016-01-11	54.218
		2016-01-12	48.265
		2016-01-13	51.852
		2016-01-14	55.662
		2016-01-15	89.844
		2016-01-16	0
		2016-01-17	54.589
		2016-01-18	50.392
		2016-01-19	50.165
		2016-01-20	57.940
		2016-01-21	51.059
		2016-01-22	90.104
		2016-01-23	0

Figura 39. Vista visitas por día

Una última perspectiva que puede ayudarnos a la toma de decisiones es ver la distribución de visitas por Zona. Viendo que zonas obtienen un mayor número de visitantes.

	CENTER	NORTH	SOUTH
Mes	Total Visitas	Total Visitas	Total Visitas
1	556.732	449.271	570.265
2	531.195	393.629	560.946
3	561.239	467.140	588.805
4	528.959	422.723	567.440
5	577.183	435.801	614.834
6	559.756	443.580	565.341
7	565.814	442.979	585.961
8	573.002	432.729	594.385
9	563.792	428.453	601.760
10	561.935	418.847	593.308
11	556.449	423.318	578.930
12	562.201	440.609	575.699

Figura 40. Total Visitas por Zona

Como podemos observar la zona Centro y Sur obtienen unos resultados similares, aunque ligeramente la zona Sur tiene mayor número de personas que van al centro comercial. La zona Norte queda alejada de los números que presentan el resto de zonas.

Una vez realizadas las Queries vamos a proceder a crear un Panel de Mando que nos permita enlazar esta información extraída del cubo Impactos-Ventas con el número de visitas efectuadas en cada centro comercial del Cubo Visitas.

7.4. Paneles de Mando.

El primer panel de mando (figura 41) muestra la relación entre las ventas-impactos y visitas mediante unos gráficos. Los gráficos muestran por Zona, Centro Comercial-Ciudad las ventas, impactos y visitas realizadas.

A partir del gráfico podemos apreciar como las visitas son más o menos constantes mientras que las ventas e impactos se ven reducidos algunos meses.

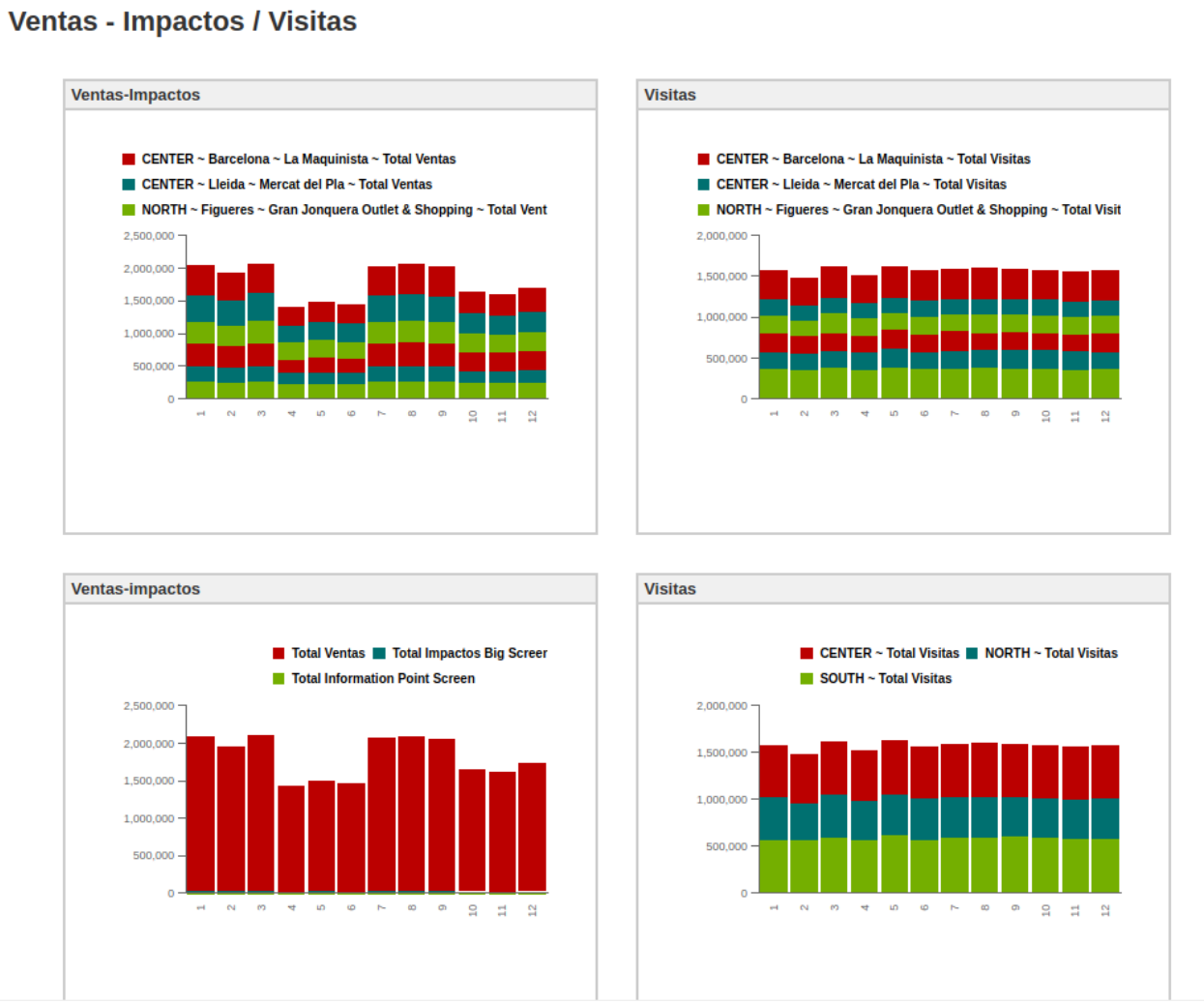


Figura 41. Panel Ventas-Impactos y Visitas

A partir de este panel no podemos decir que hay una influencia entre el número de visitantes y el número de impactos o ventas. No obstante, para estar más

seguros se ha creado un nuevo panel que muestre únicamente los totales por mes entre el número de Ventas y el número de Visitantes para poder determinar si hay alguna relación entre ambos. Podemos obtener una vista más ampliada de la figura 41 revisando el Anexo 13 de este documento.

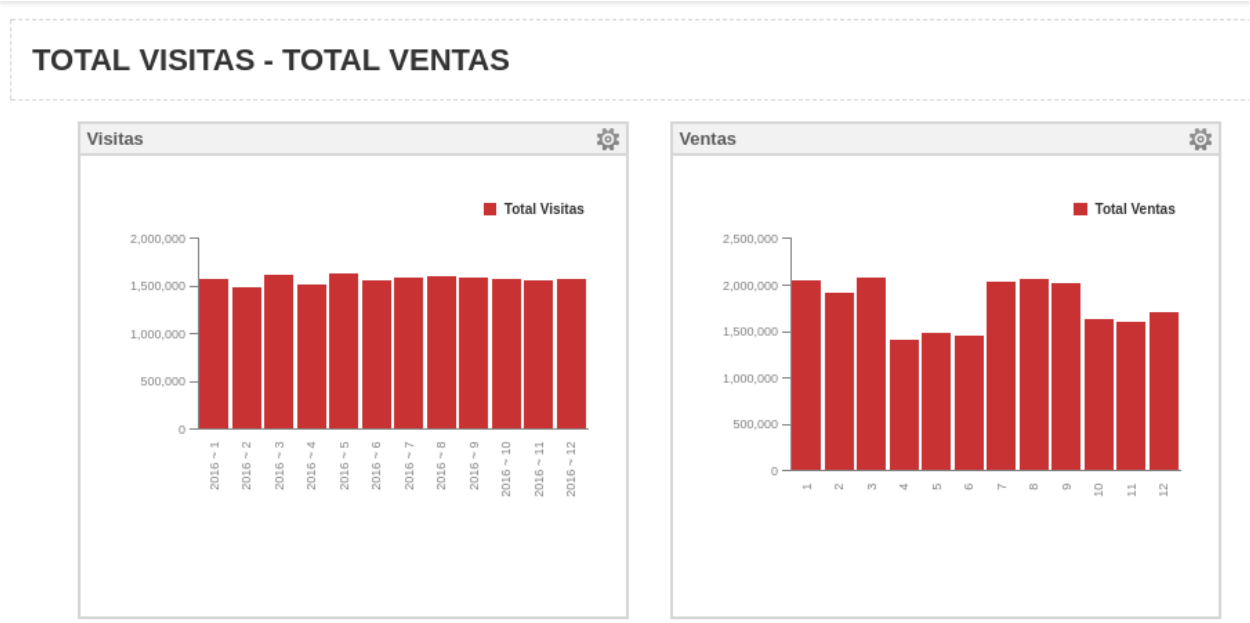


Figura 42. Panel Total Visitas – Total Ventas

Como podemos observar, se confirma lo que anteriormente se ha comentado, el número de visitantes es más o menos constante, siendo incluso ligeramente superior el mes de Mayo, sin embargo las ventas varían dependiendo de la época del año, sin ir más lejos el mes de mayo es uno de los meses de menor número de ventas.

Podemos concluir, que no hay una relación entre el número de visitantes y el número de ventas.

7.5. Conclusiones de la campaña publicitaria.

El último punto de todo proceso de Business Intelligence es el análisis de los datos para sacar conclusiones. En nuestro proyecto una vez realizadas las consultas podemos determinar el grado de efectividad de la campaña de publicitaria basada en Impactos.

En lo que respecta al problema planteado en el presente TFM, el presente punto se basará en dar una respuesta a la cuestión planteada en el punto anterior *"En general, ¿la publicidad basada en impactos visuales, es efectiva?"*

En líneas generales tras responder a las cuestiones planteadas mostradas en el punto anterior podríamos decir que si, *SI* hay un grado de efectividad entre la relación de impactos y las ventas efectuadas en cada caso, *"los números relfejan que cuantos más impactos, mayor número de ventas"*. Aunque también añadiría a la respuesta que daría al gerente de una empresa un "depende". Depende de la zona, el producto, el momento del año…
Por ejemplo, si estas en la zona Centro y tienes una empresa de aparatos Electrónicos te interesa que se realice esta campaña publicitaria dado que tus ventas si incrementarán notablemente.

No obstante, sería interesante disponer de datos de ventas de otros años en los que no se hubiera efectuado esta campaña publicitaria para poder comparar el número de Ventas y poder sacar conclusiones más acertadas.

En resumen, las conclusiones que podríamos extraer del análisis realizado son las siguientes:

- ✓ **Los números reflejan que a mayor número de impactos, mayor número de Ventas.**

- ✓ **Con un número similar de impactos, la Zona Centro muestra una mayor efectividad obteniendo un mayor número de ventas con respecto a la zona Norte o Sur.**

- ✓ **Los artículos de la familia de productos "Electronics" son los que mayor eficacia presentan en líneas generales.**

- ✓ **El primer trimestre del año y los meses de verano presentan una mayor eficacia, siendo los meses de más impactos y más ventas.**

- ✓ **Existe una eficacia en la campaña dependiendo de la zona y el producto.**

- ✓ **En líneas generales la eficacia de la campaña es la misma en todas las zonas.**

- ✓ **Los números reflejan una afluencia de visitantes constante durante todo el año. Esto no influye en el número de ventas o impactos.**

8. Conclusiones

Antes de realizar este TFM no había visto nada en mi vida profesional relacionado con herramientas BI, salvo trabajos artesanales hechos mediante Hojas de Cálculo y macros. La realización de este TFM relacionado con BI despertó mi interés debido a la amplia demanda de especialistas en este sector y al bajo número de personas con conocimientos en la materia. No obstante, al margen de esta creciente demanda tras realizar el proyecto pienso que me va a ser de mucha ayuda en mi vida laboral ya que el disponer del conocimiento para poder explotar los datos para sacar conclusiones, me ayudarán en futuras tomas de decisiones.

Las lecciones aprendidas que me llevo de este proyecto son muchas ya que todo el camino ha sido nuevo para mí. Destacaré la importancia de diseñar bien el DWH y tener conocimientos en ETL. ETL a pesar de parecer un proceso sencillo gracias a herramientas como el Spoon de Pentaho, es el que ha llevado más tiempo. No menos importante, es tener claro cómo será el cubo con el que vamos a trabajar ya que esto determinará el diseño de nuestro DWH.

En líneas generales, en mi proyecto se ha seguido la planificación tal y como se planteó desde el primer día. Por otro lado, la metodología elegida también ha sido la correcta. Dado que Scrum es una metodología ágil de gestión de proyectos cuyo objetivo primordial es elevar al máximo la productividad de un equipo. Reduce al máximo la burocracia, y se centra en producir software que funcione y en ofrecer resultados visibles en poco tiempo, incluso ante requerimientos inestables, en mi caso ha sido lo que he hecho. Mi objetivo siempre ha sido producir un “entregable” que permitiera dar valor inmediato al trabajo realizado, por esta razón lo primero que hice fue centrarme en el Cubo más sencillo, el cubo Visitas. Una vez conseguí, diseñarlo e implementarlo, comencé la creación del cubo Impactos – Ventas. Esto me ha permitido terminar el TFM a tiempo y con una calidad aceptable.

Otro punto a considerar es la experiencia y conocimientos. Para proyectos BI considero muy importante la experiencia que tengas ya no solo en el manejo de este tipo de herramientas, sino también los conocimientos que tengas sobre la materia que estás tratando en el proyecto BI.

En lo que respecta a las líneas a explorar en el futuro que no se han podido desarrollar en este trabajo, he de comentar que estoy planteando la posibilidad de realizar un master de Business Intelligence que me permita ver alternativas a Pentaho en el mercado y técnicas de minería de datos. Por otro lado, y dada la falta de tiempo, echo de menos el poder haber realizado un análisis más exhaustivo en el caso de haber dispuesto de más información.

En resumen, considero que ha sido una grata experiencia la realización de este TFM cuyos conocimientos voy a utilizar, con mucha probabilidad, en mi futuro profesional.

9. Glosario

- **Business**: Término en inglés que significa negocio.

- **Intelligence**: Término en inglés que significa Inteligencia.

- **BI**: Siglas acuñadas del inglés que significan Business Intelligence, en castellano, Inteligencia de Negocio.

- **ETL**: Extract, Transform and Load. Siglas para referirnos al proceso de Extracción, transformación y carga de los datos en bruto al datawarehouse.

- **OLAP**: OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia de negocios (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

- **Cubo**: Un cubo OLAP es una estructura de datos que supera las limitaciones de las bases de datos relacionales y proporciona un análisis rápido de datos. Los cubos pueden mostrar y sumar grandes cantidades de datos, a la vez que proporcionan a los usuarios acceso mediante búsqueda a los puntos de datos.

- **Scrum**: Scrum es un proceso en el que se aplican de manera regular un conjunto de buenas prácticas para trabajar colaborativamente, en equipo, y obtener el mejor resultado posible de un proyecto.

- **Datawarehouse**: Un Datawarehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de pespectivas y con grandes velocidades de respuesta.

- **Schema**: Es habitual referirse a los esquemas como "XML schema" de forma genérica, pero se recomienda utilizar el término “documento esquema” (schema document) o "definición de esquema"(schema definition), y reservar “XML Schema” para la denominación de este lenguaje específico.

- **Query**: Query string o, en español, cadena de consulta es un término informático que se utiliza para hacer referencia a una interacción con una base de datos. Es la parte de una URL que contiene los datos que deben pasar a aplicaciones web como los programas CGI.

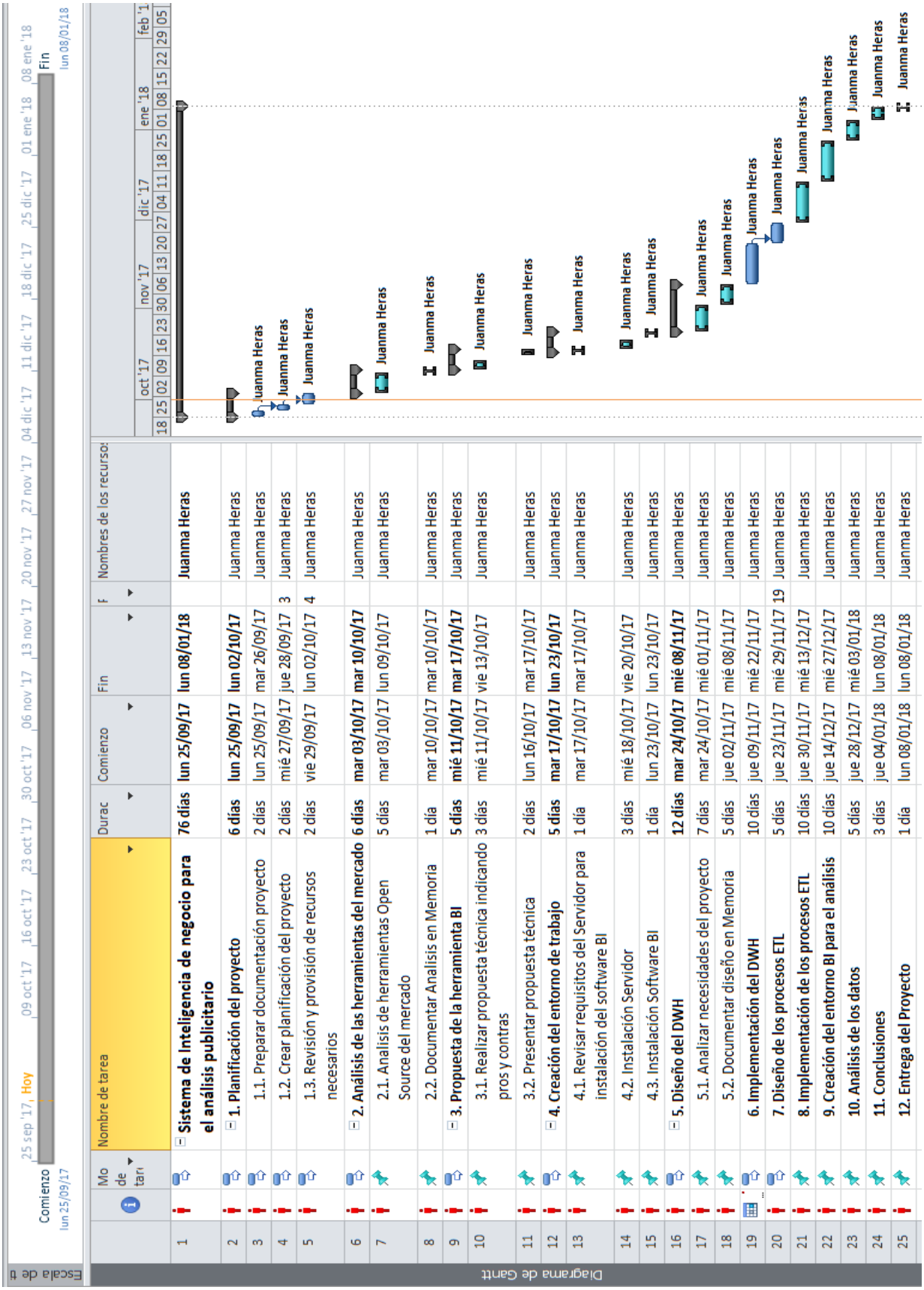
- **DWH**: Siglas que hacen referencia a Datawarehouse.

10. Bibliografía

[1] https://es.wikipedia.org/wiki/Scrum_(desarrollo_de_software)
[2] https://www.businessintelligence.info/definiciones/scrum-business-intelligence.html
[3] https://blog.capterra.com/top-8-free-and-open-source-business-intelligence-software/
[4] http://www.stratebi.com/spagobi
[5] https://es.wikipedia.org/wiki/Pentaho
[6] https://help.pentaho.com/Documentation/5.2/0D0/160/000
[7] https://blog.jaapjansma.nl/2016/05/11/installing-pentaho-6-1-community-edition-on-ubuntu/
[8] http://mundodb.es/introduccion-a-los-datawarehouse
[9] https://es.wikipedia.org/wiki/Tabla_de_dimensi%C3%B3n
[10] https://es.wikipedia.org/wiki/Tabla_de_hechos
[11] https://es.wikipedia.org/wiki/Esquema_en_copo_de_nieve
[12] https://churriwifi.wordpress.com/2010/05/01/16-procesos-etl-escenarios-para-el-diseno-de-los-procesos/

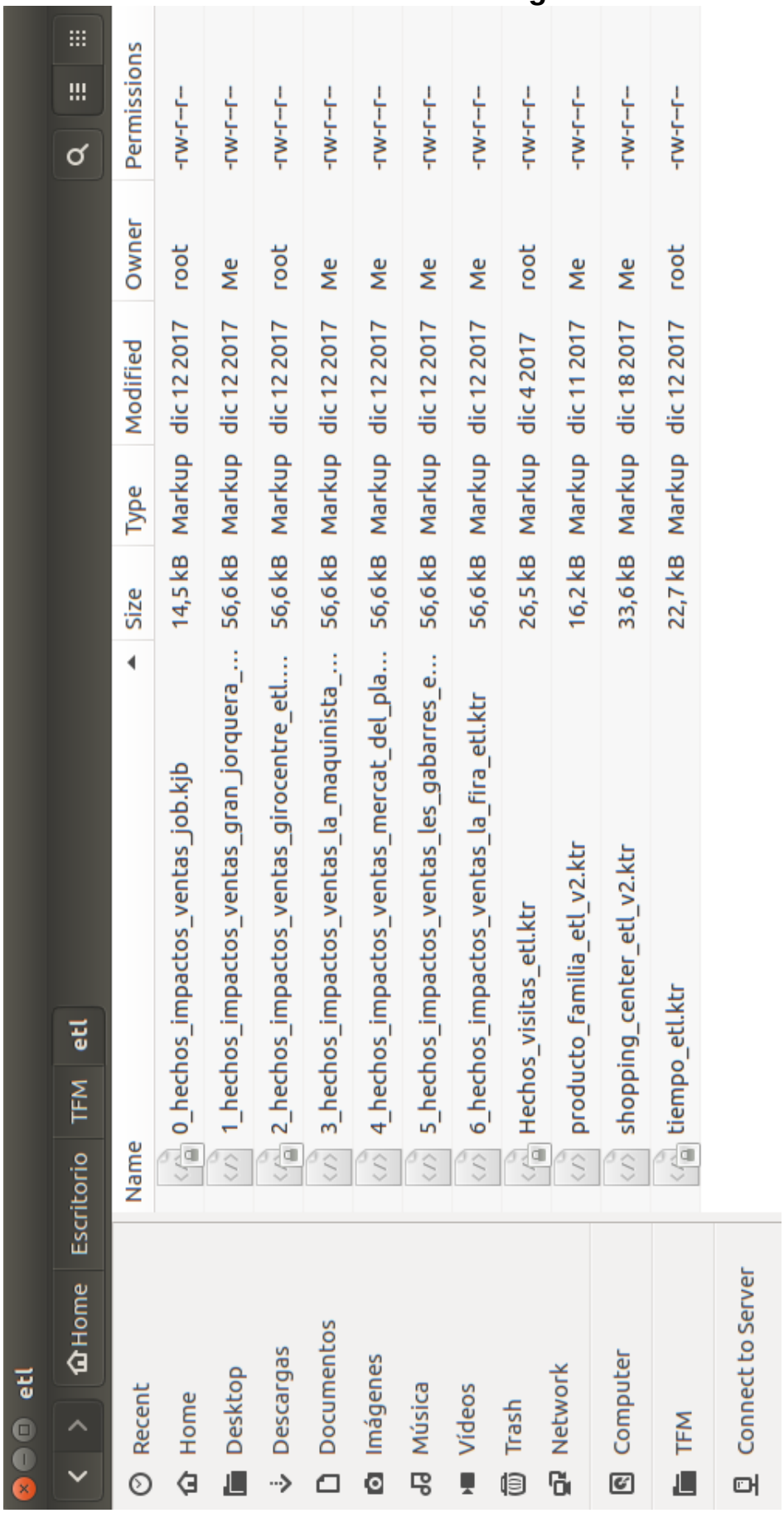
11. Anexos

Anexo 1: Diagrama de Gantt

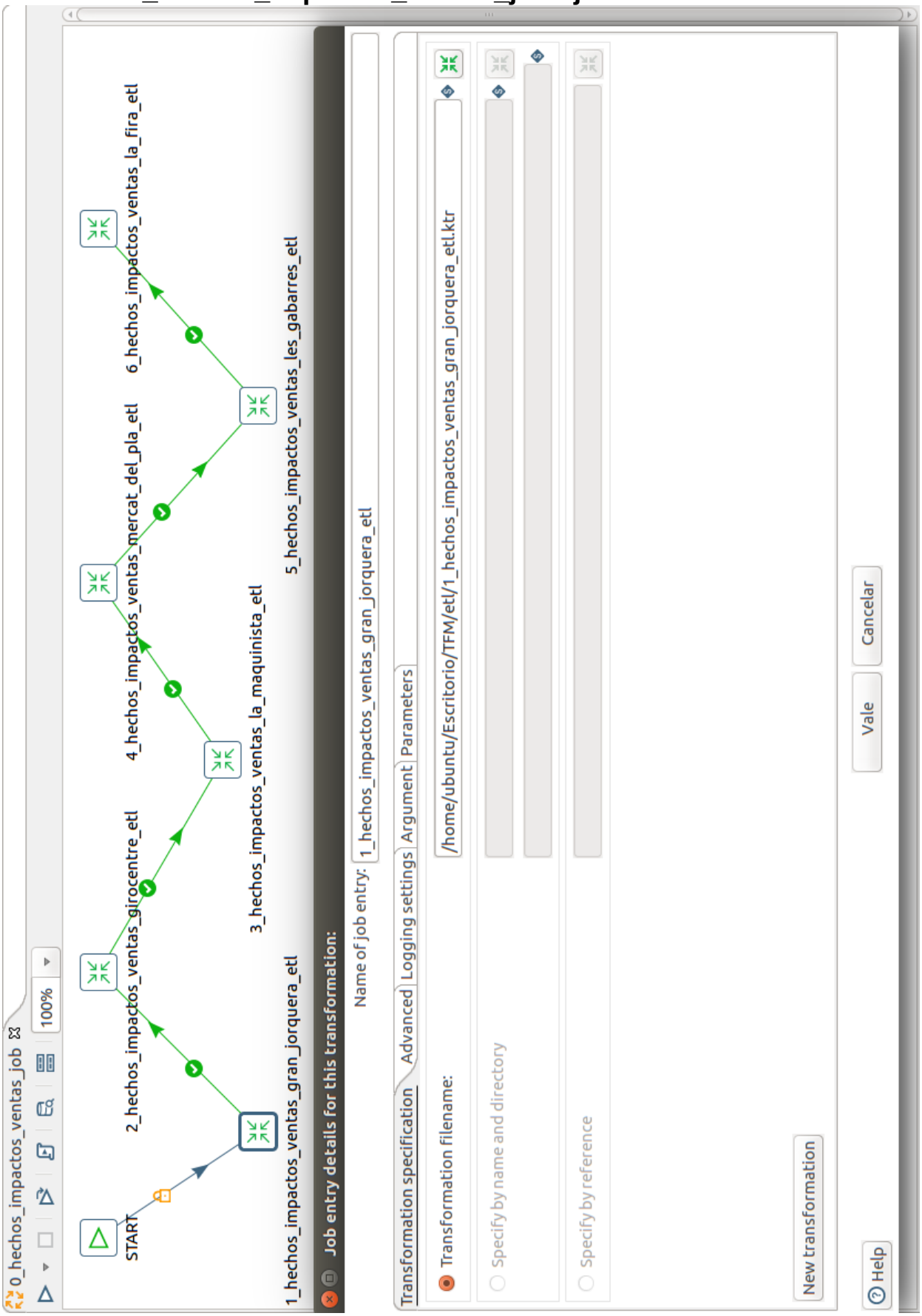


	Mo de tar	Nombre de tarea	Durac	Comienzo	Fin	F	Nombres de los recurso
1		**Sistema de Inteligencia de negocio para el análisis publicitario**	**76 días**	**lun 25/09/17**	**lun 08/01/18**		**Juanma Heras**
2		**1. Planificación del proyecto**	**6 días**	**lun 25/09/17**	**lun 02/10/17**		Juanma Heras
3		1.1. Preparar documentación proyecto	2 días	lun 25/09/17	mar 26/09/17		Juanma Heras
4		1.2. Crear planificación del proyecto	2 días	mié 27/09/17	jue 28/09/17	3	Juanma Heras
5		1.3. Revisión y provisión de recursos necesarios	2 días	vie 29/09/17	lun 02/10/17	4	Juanma Heras
6		**2. Análisis de las herramientas del mercado**	**6 días**	**mar 03/10/17**	**mar 10/10/17**		Juanma Heras
7		2.1. Análisis de herramientas Open Source del mercado	5 días	mar 03/10/17	lun 09/10/17		Juanma Heras
8		2.2. Documentar Analisis en Memoria	1 día	mar 10/10/17	mar 10/10/17		Juanma Heras
9		**3. Propuesta de la herramienta BI**	**5 días**	**mié 11/10/17**	**mar 17/10/17**		Juanma Heras
10		3.1. Realizar propuesta técnica indicando pros y contras	3 días	mié 11/10/17	vie 13/10/17		Juanma Heras
11		3.2. Presentar propuesta técnica	2 días	lun 16/10/17	mar 17/10/17		Juanma Heras
12		**4. Creación del entorno de trabajo**	**5 días**	**mar 17/10/17**	**lun 23/10/17**		Juanma Heras
13		4.1. Revisar requisitos del Servidor para instalación del software BI	1 día	mar 17/10/17	mar 17/10/17		Juanma Heras
14		4.2. Instalación Servidor	3 días	mié 18/10/17	vie 20/10/17		Juanma Heras
15		4.3. Instalación Software BI	1 día	lun 23/10/17	lun 23/10/17		Juanma Heras
16		**5. Diseño del DWH**	**12 días**	**mar 24/10/17**	**mié 08/11/17**		Juanma Heras
17		5.1. Analizar necesidades del proyecto	7 días	mar 24/10/17	mié 01/11/17		Juanma Heras
18		5.2. Documentar diseño en Memoria	5 días	jue 02/11/17	mié 08/11/17		Juanma Heras
19		**6. Implementación del DWH**	10 días	jue 09/11/17	mié 22/11/17		Juanma Heras
20		**7. Diseño de los procesos ETL**	5 días	jue 23/11/17	mié 29/11/17	19	Juanma Heras
21		**8. Implementación de los procesos ETL**	10 días	jue 30/11/17	mié 13/12/17		Juanma Heras
22		**9. Creación del entorno BI para el análisis**	10 días	jue 14/12/17	mié 27/12/17		Juanma Heras
23		**10. Análisis de los datos**	5 días	jue 28/12/17	mié 03/01/18		Juanma Heras
24		**11. Conclusiones**	3 días	jue 04/01/18	lun 08/01/18		Juanma Heras
25		**12. Entrega del Proyecto**	1 día	lun 08/01/18	lun 08/01/18		Juanma Heras

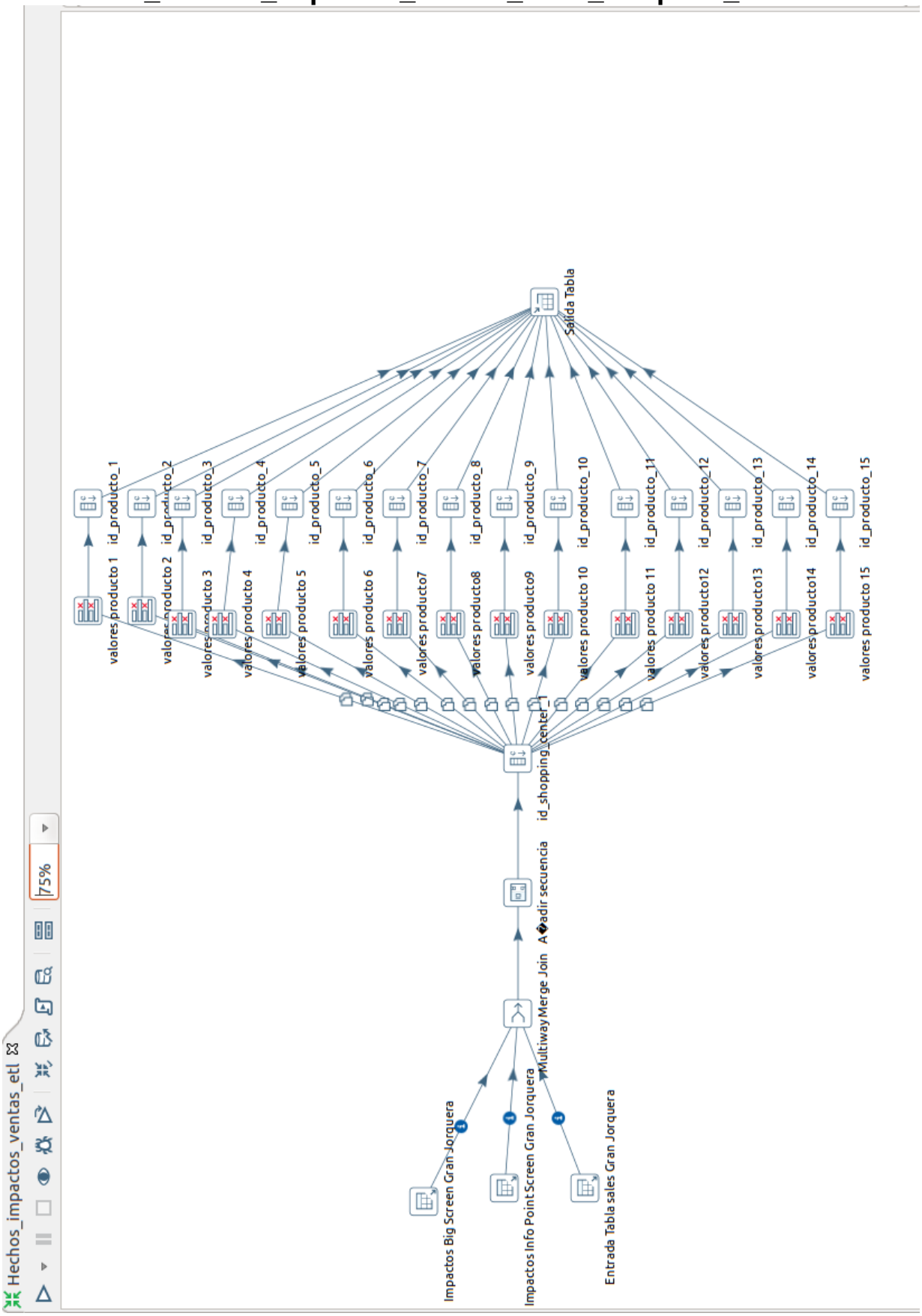
Anexo 2: Relación de archivos ETL generados



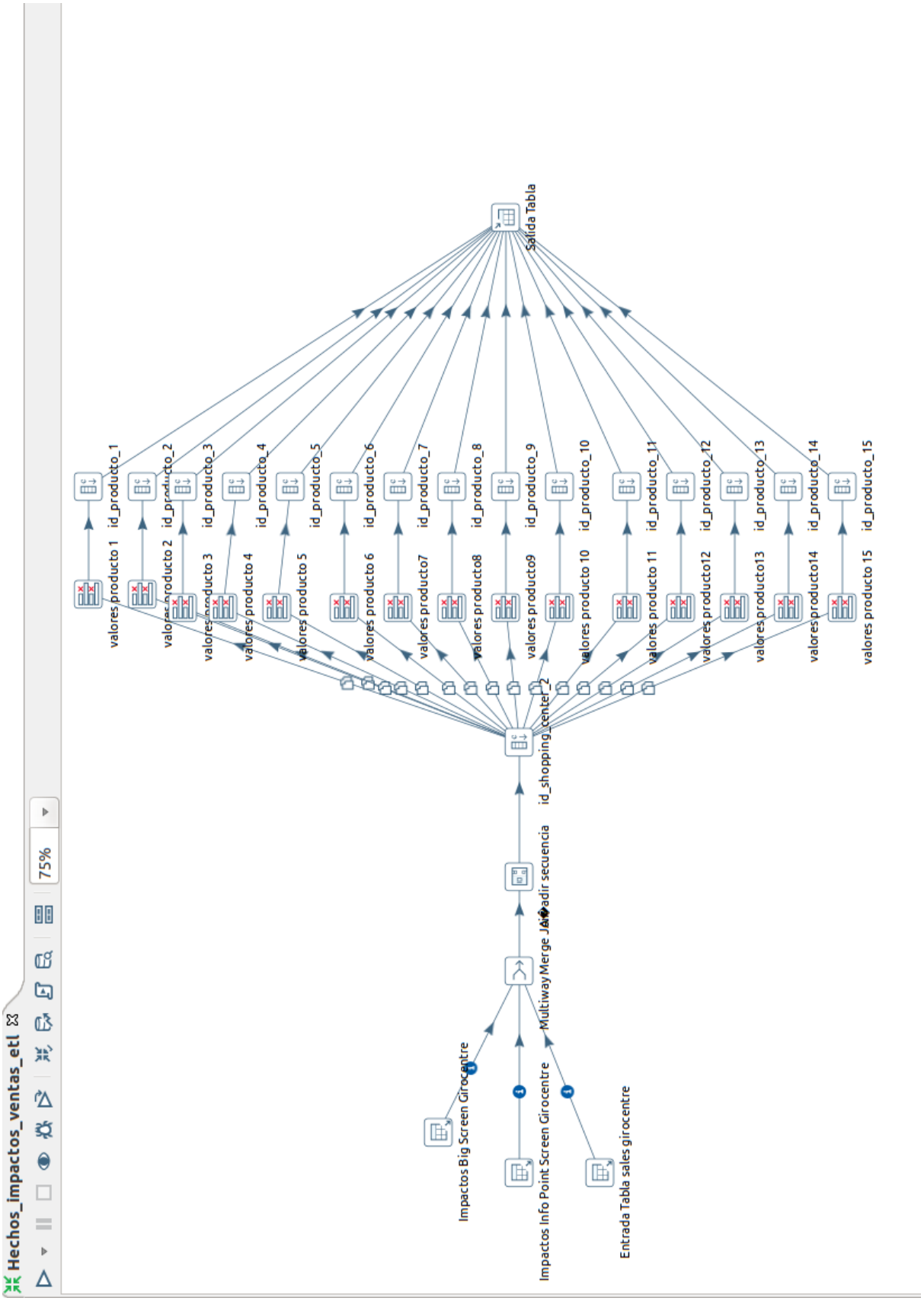
Anexo 3: 0_hechos_impactos_ventas_job.kjb



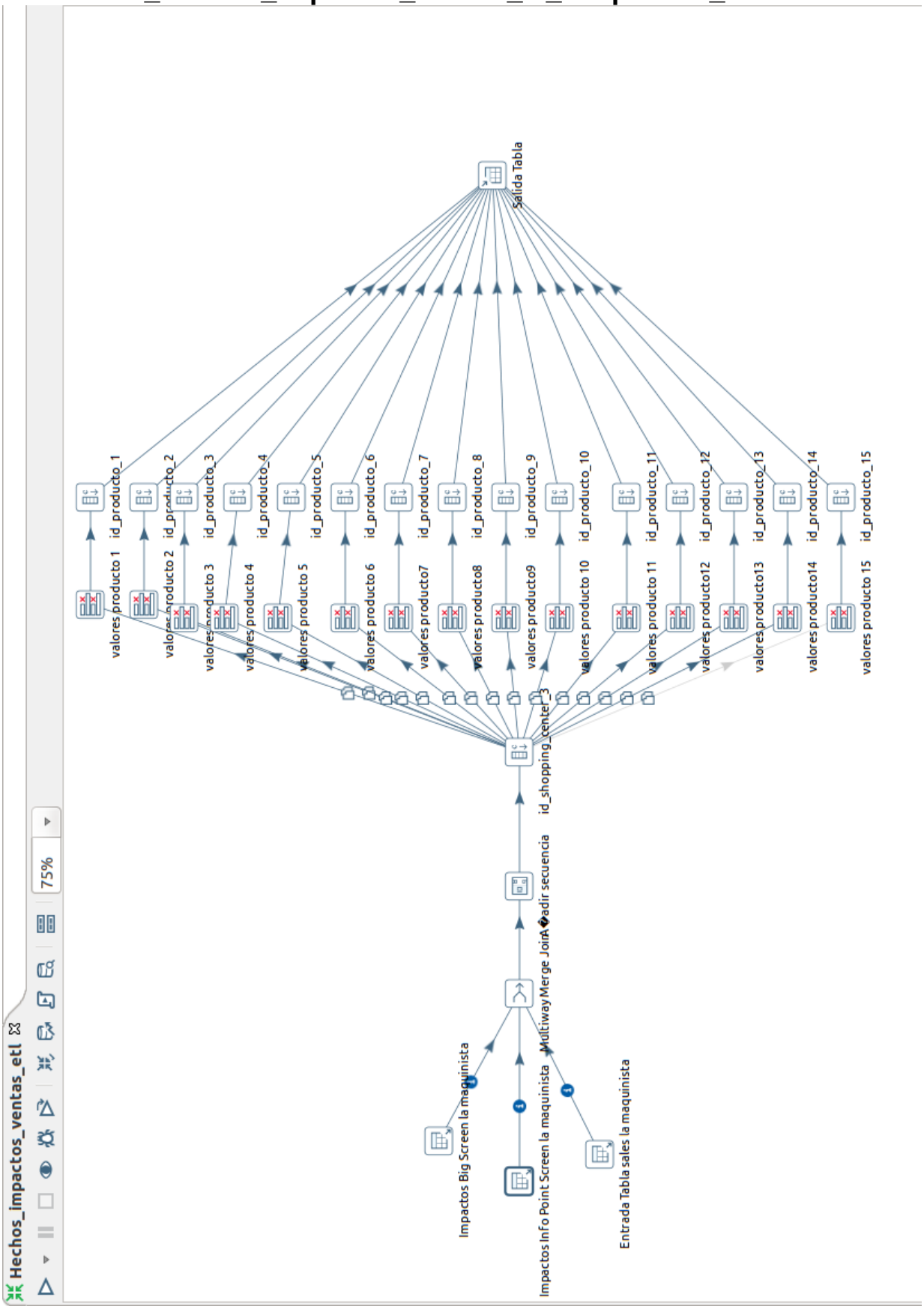
Anexo 3: 1_Hechos_Impactos_Ventas_Gran_Jorquera_etl.ktr



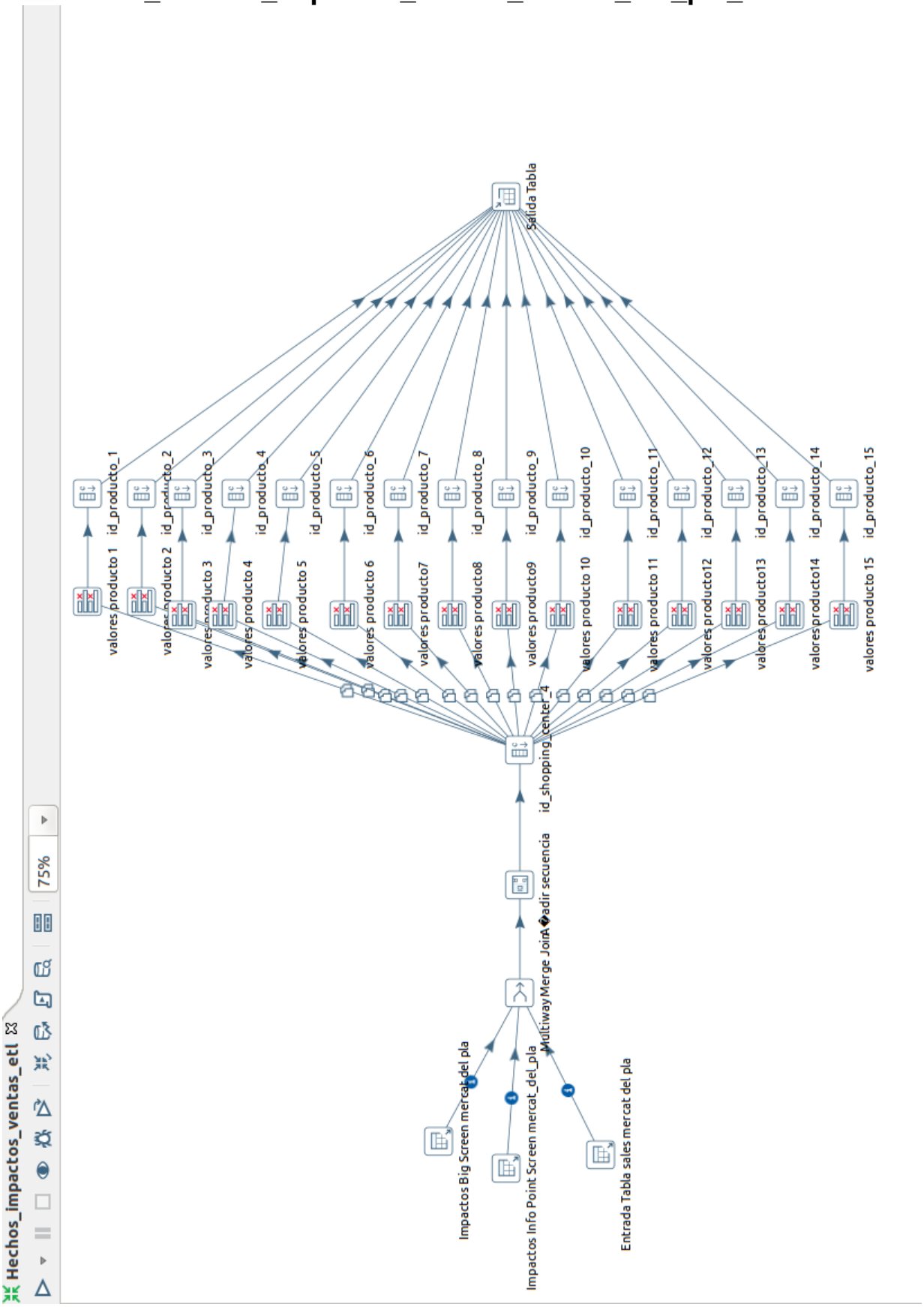
Anexo 4: 2_Hechos_Impactos_Ventas_Girocentre_etl.ktr



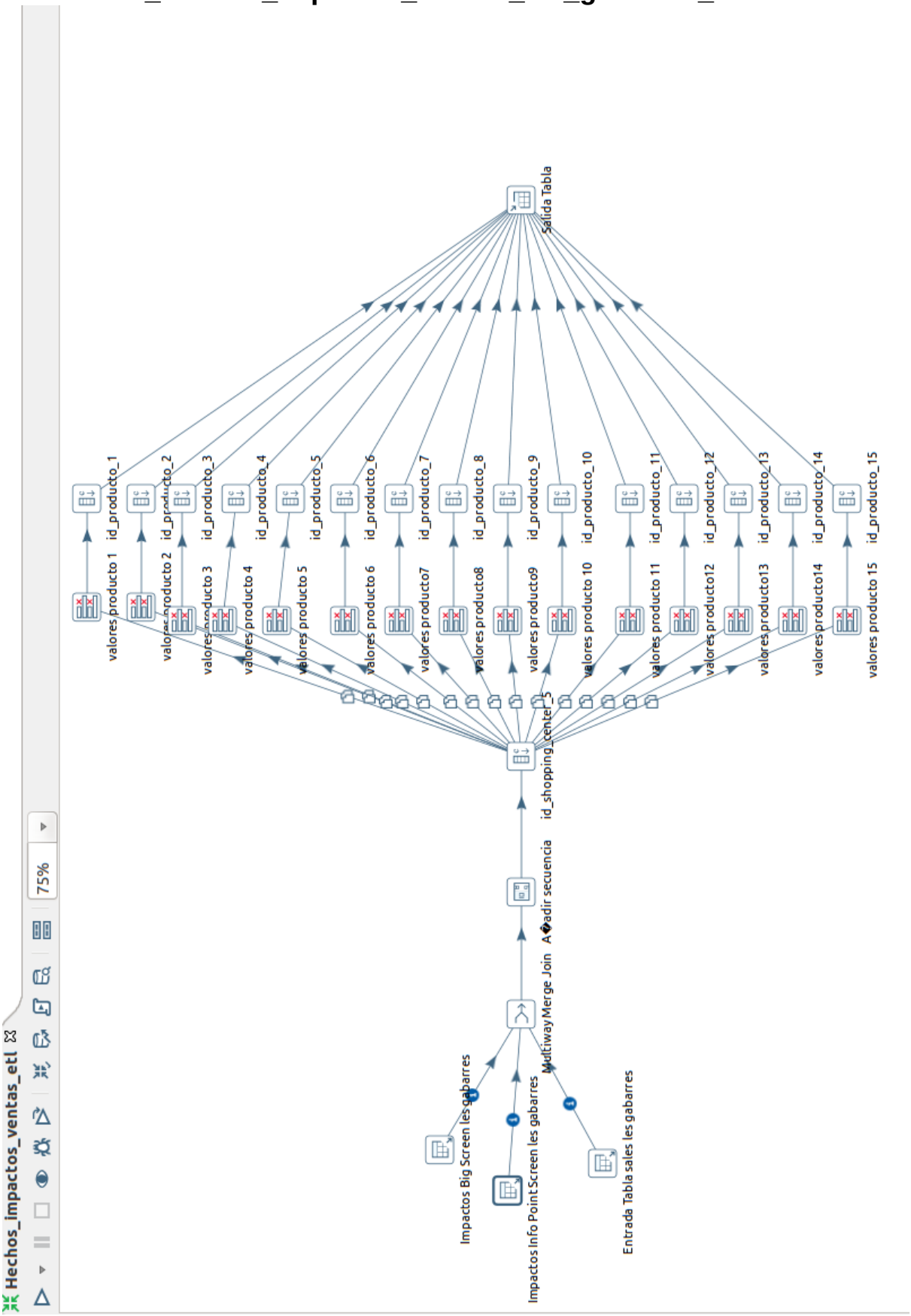
Anexo 5: 3_Hechos_Impactos_Ventas_la_maquinista_etl.ktr



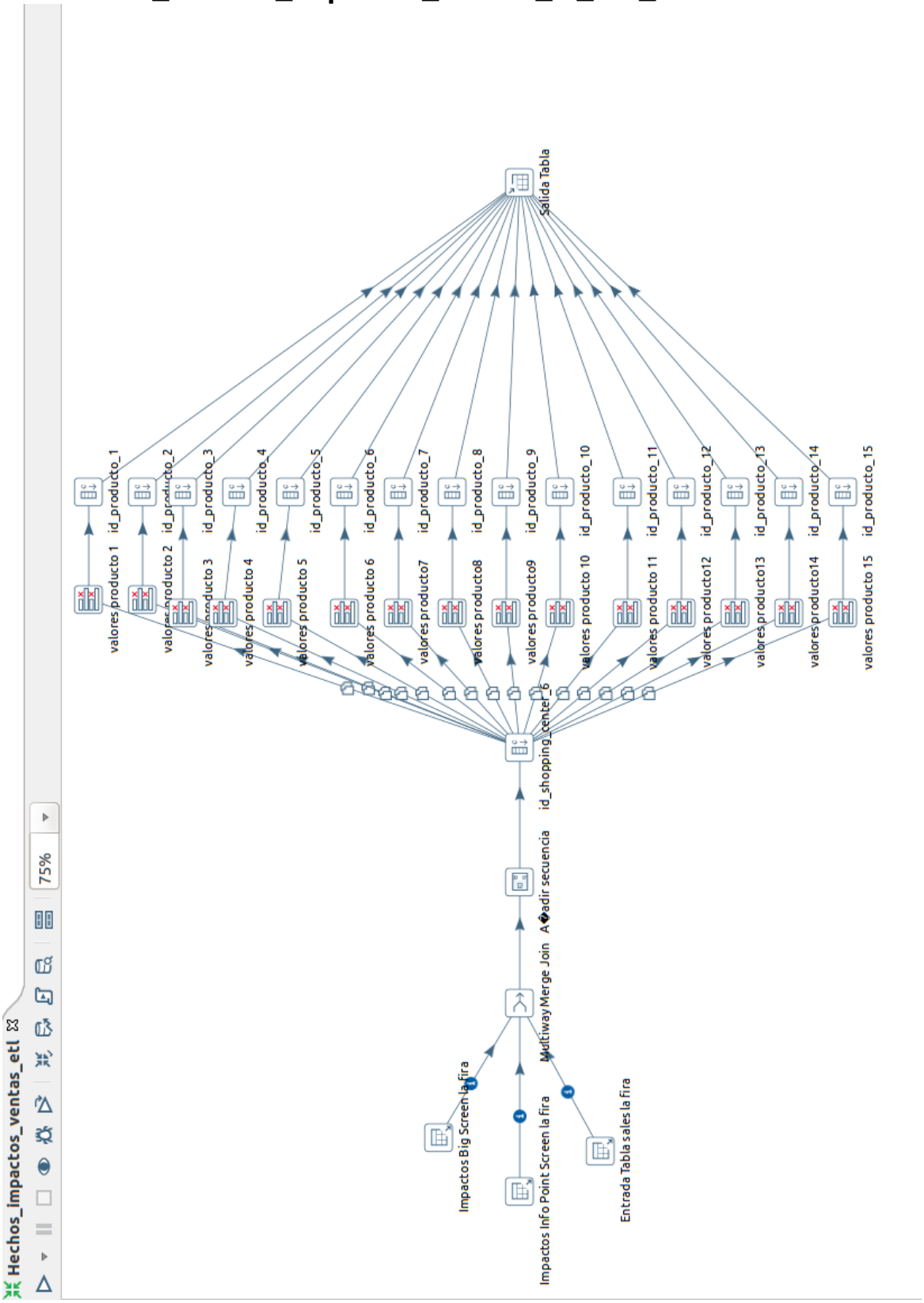
Anexo 6: 4_Hechos_Impactos_Ventas_mercat_del_pla_etl.ktr



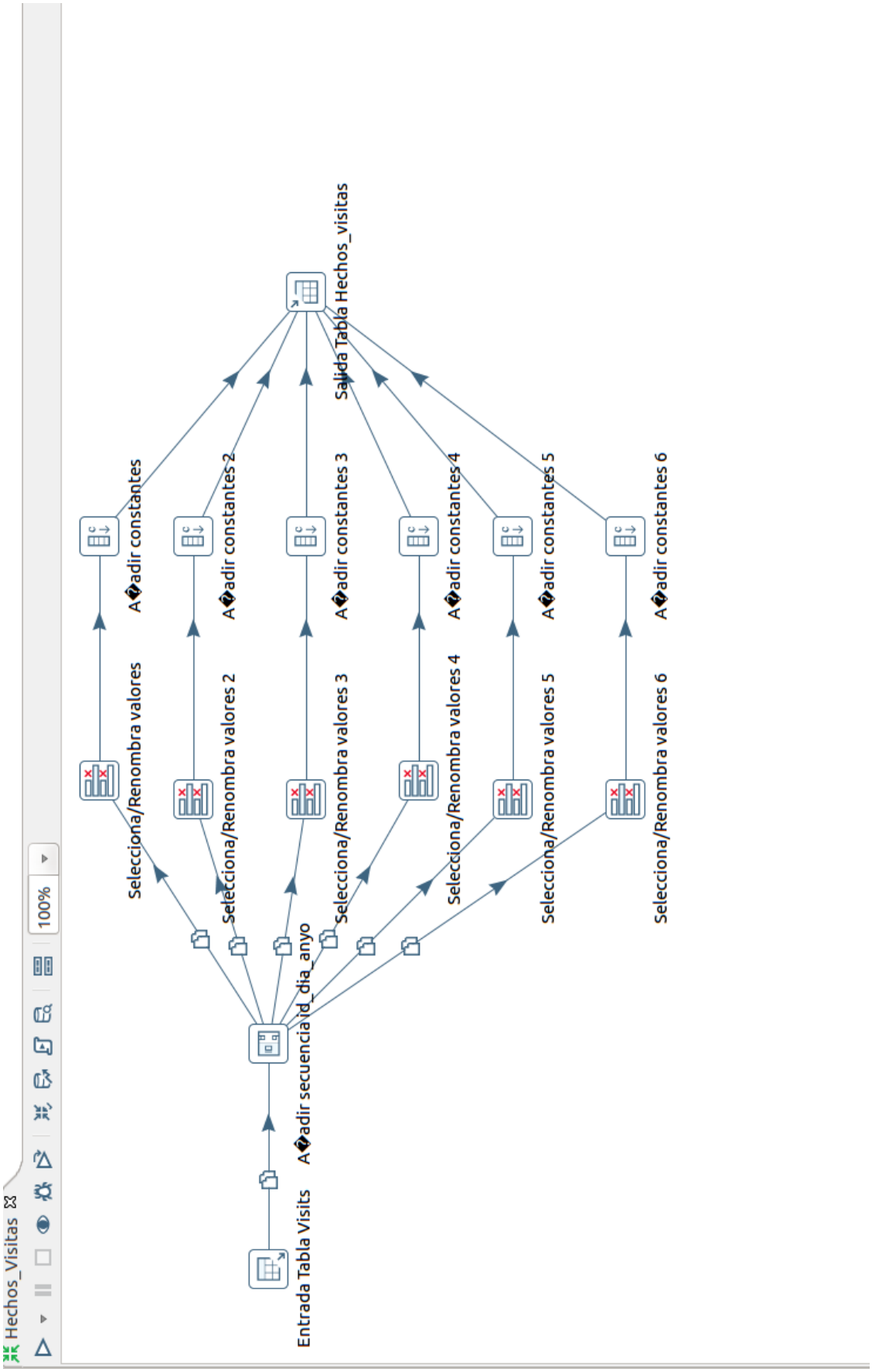
Anexo 7: 5_Hechos_Impactos_Ventas_les_gabarres_etl.ktr



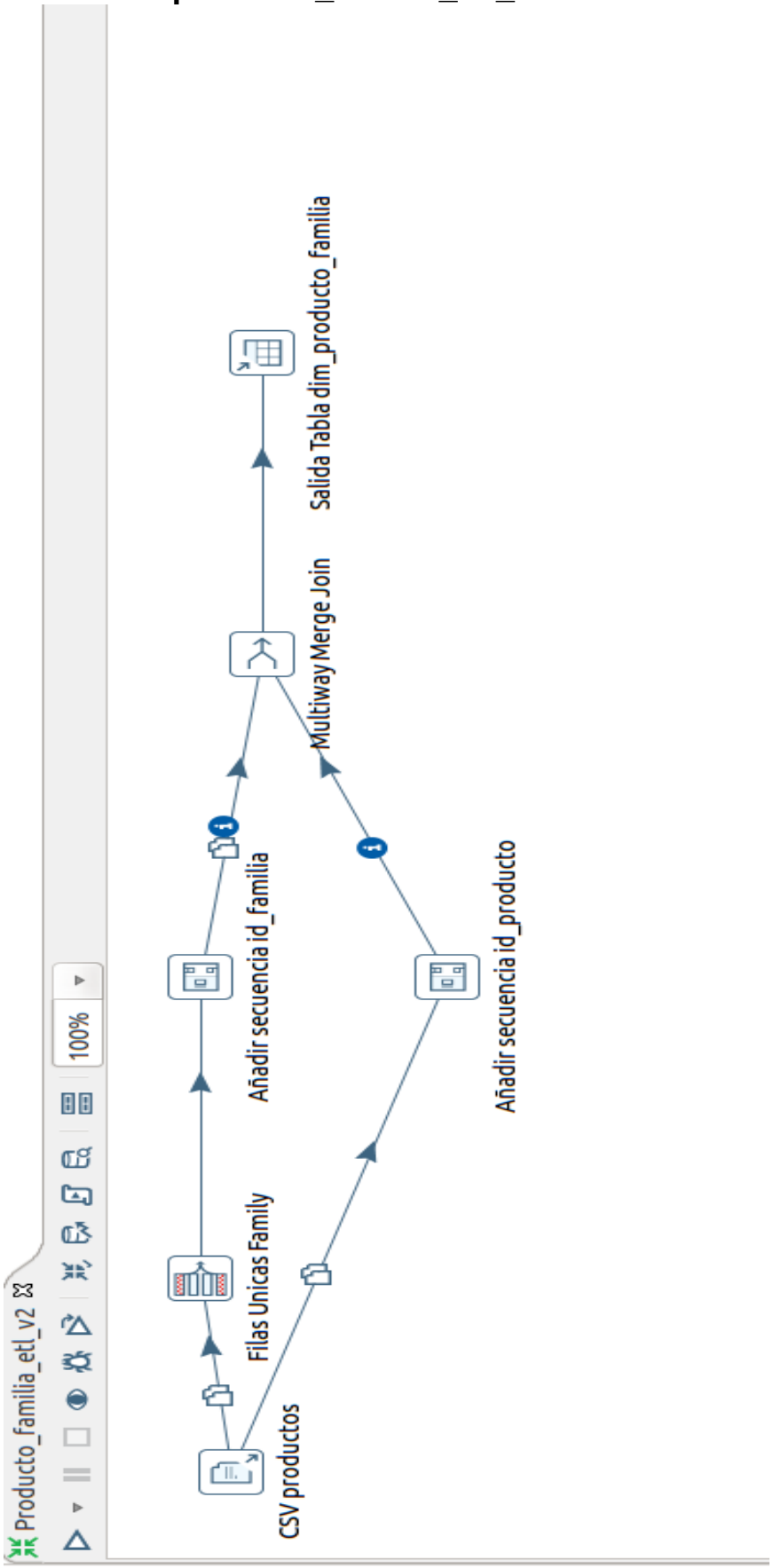
Anexo 8: 6_Hechos_Impactos_Ventas_la_fira_etl.ktr



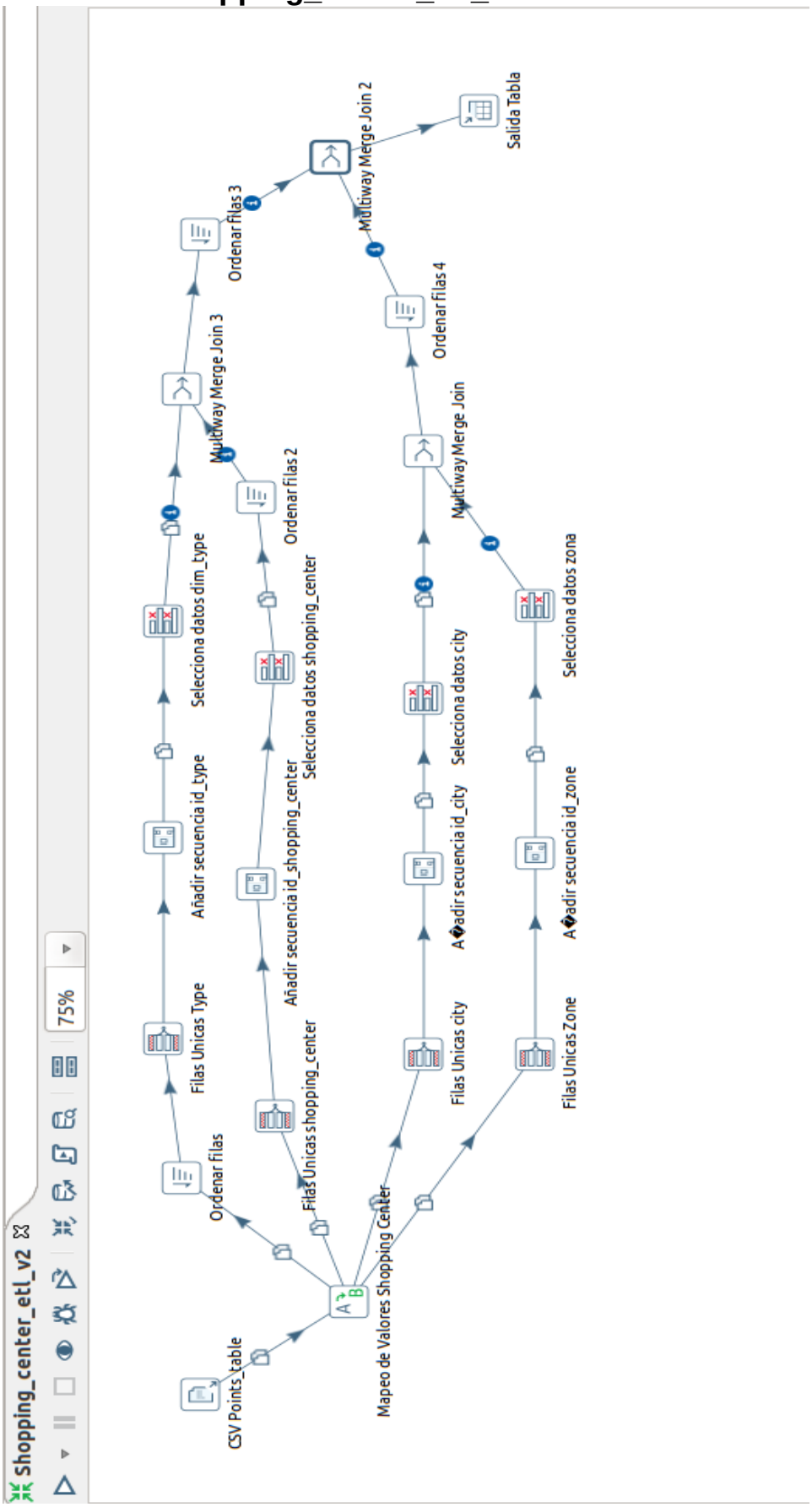
Anexo 9: Hechos_Visitas_etl.ktr



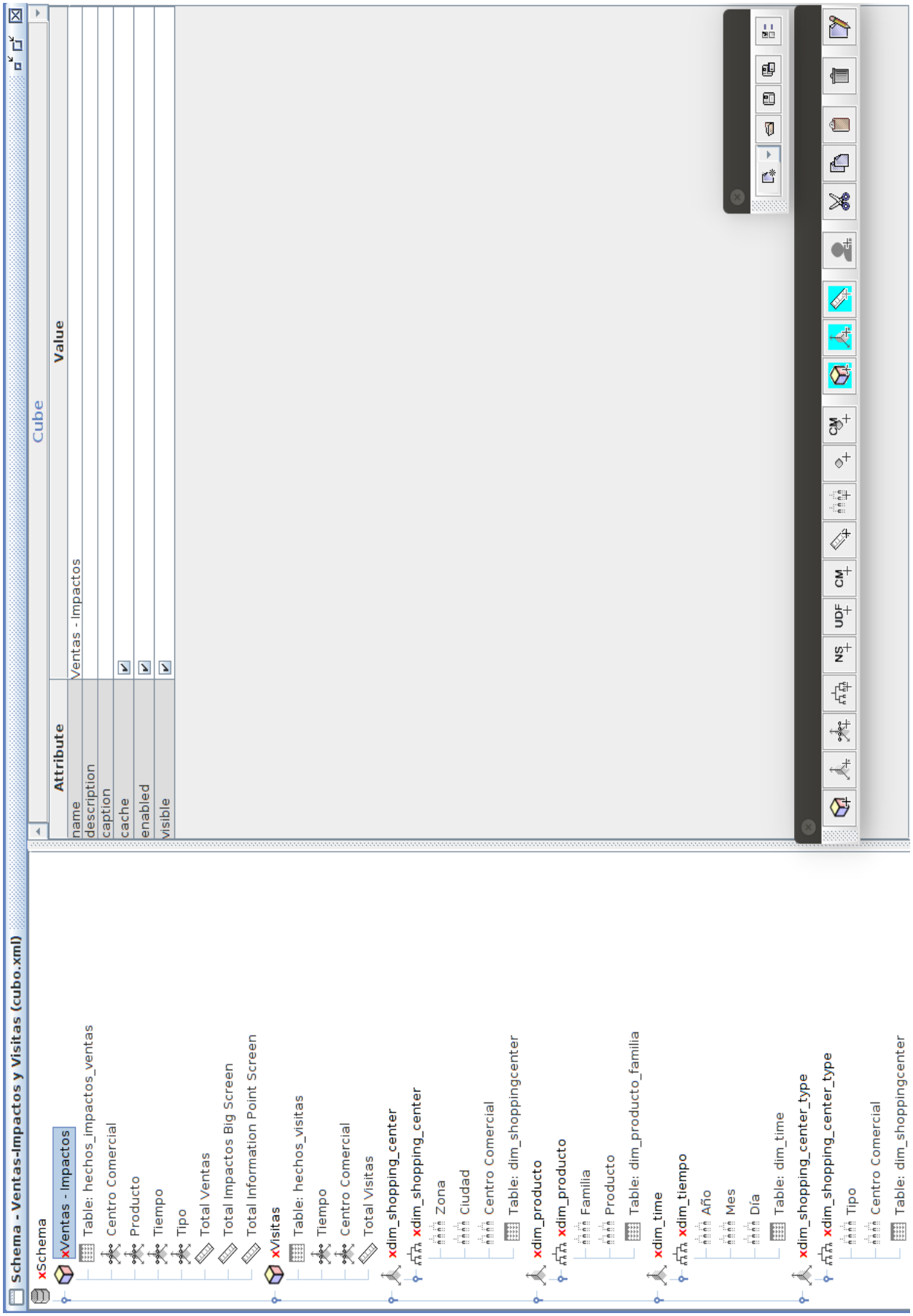
Anexo 10: producto_familia_etl_v2.ktr



Anexo 11: shopping_center_etl_v2.ktr



Anexo 12: cubo.xml



Anexo 13: Panel Ventas – Impactos