Grado de Ingeniería Informática

Proyecto BI para el análisis del proceso de captación de alumnos en centros educativos.

Trabajo Fin de Grado

Estudiante
David Ropero Alcázar

Consultoría
Xavier Martínez Fonte

Fecha
Enero 2021

Resumen del trabajo (máximo 250 palabras)

Título	Proyecto BI para el análisis del proceso de captación de alumnos en centros educativos.
Autor	David Ropero Alcázar
Consultor	Xavier Martínez Fonte
Fecha de entrega	07/01/2021
Titulación	Grado en Ingeniería Informática
Área de especialización del proyecto	Business Intelligence
Palabras clave	Modelado de base de datos, Data Warehouse, ETL, Dashboard, Marketing.

El presente trabajo fin de grado tiene como objetivo recorrer todas las fases que un departamento de marketing necesita para la consecución de uno de sus principales objetivos: captar clientes. Para ello debe conocer el perfil de su cliente tipo, las preferencias sobre los productos o servicios y, finalmente, cuantificar la conversión que suponen sus acciones de marketing. Concretamente se ha elegido el sector educativo y la captación de alumnos potenciales.

Este proceso necesita superar varios pasos intermedios en cualquier proyecto BI: modelar de forma óptima el almacén de datos, realizar cargas de datos iniciales y utilizar herramientas de análisis BI que permitan extraer información y conocimiento de forma rápida para el rol del director de marketing. Durante todo este proceso se han aplicado técnicas de gestión de proyectos como la planificación con diagramas de Gantt y el seguimiento de tareas mediante tableros Kanban.

El proyecto ha simulado los datos principales que dan contenido al *data warehouse* y se ha apoyado de forma intensiva en la potencia del lenguaje SQL como principal herramienta para extraer la información. Otro objetivo ha sido usar gráficos no tan comunes en el entorno empresarial que, sin embargo, están ahí para asistir de forma visual en el caso de información compleja como pueden ser los flujos de datos entre diferentes momentos del tiempo.

Finalmente se pone a disposición un entregable de los diferentes *dashboard* así como de las conclusiones que plantearán posibles líneas de actuación.

Abstract (in English, 250 words or less)

The present end-of-degree work aims to cover all the phases that a marketing department needs to achieve one of its main objectives: to capture customers. To do so, it must know the profile of its typical customer, the preferences about products or services and, finally, quantify the conversion that its marketing actions involve. Specifically, the educational sector and the recruitment of potential students have been chosen.

This process needs to go through several intermediate steps in any BI project: optimally model the data warehouse, perform initial data loads and use BI analysis tools that allow to extract information and knowledge quickly for the role of the marketing director. This path has been leveraged to use project management techniques such as planning with Gantt charts and tracking tasks using Kanban dashboards.

The project has simulated the main data that give content to the data warehouse and has relied heavily on the power of SQL language as the main tool for extracting information. Another objective has been to use graphics that are not so common in the business environment but that are there to visually assist complex information such as data flows between different moments in time.

Finally, a deliverable of the different dashboards is made available, as well as the conclusions that will raise possible lines of action.

Índice de contenidos

Capítulo 1 – Contextualizando el proyecto

Capítulo 2 – Definiendo la base conceptual

Capítulo 3 – Construcción del Data Warahouse

Capítulo 4 – Visualización en Power BI

Capítulo 1 – Contextualizando el proyecto

1.1 Introducción

El marketing ha evolucionado en las últimas décadas gracias a los grandes avances en computación y, especialmente, a la ingente cantidad de datos que se ha podido recabar de los hábitos de consumo de los usuarios. En el área analítica, son muchas las técnicas estadísticas que se han afinado y adaptado para servir a los dos principales objetivos de cualquier departamento de marketing: captar clientes y fidelizar a los clientes actuales.

Todo ha derivado en especializar este área del marketing en lo que se conoce como Inteligencia de Negocio. El principal cometido es transformar los datos en información y la información en conocimiento. Finalmente ese conocimiento, asistido de herramientas gráficas y predictivas, permite tomar las decisiones más idóneas para alcanzar los dos objetivos anteriormente citados.

Sin embargo, la inteligencia de negocio no es una ciencia exacta. En cada fase del proceso, las opciones que se abren son tan amplias que es imprescindible acotar datos e información para poder obtener conclusiones que se puedan contrastar con los resultados, evitando caer en correlaciones causales tanto si los objetivos se cumplen como si no. No existe un modelo perfecto y mucho menos una herramienta perfecta, sino que requiere de sucesivas iteraciones y una constante evolución y adaptación a los cambios que se producen en el consumidor y en el mercado así como la interacción entre diferentes herramientas y utilidades concretas a cada fase y objetivo.

Para elaborar el modelo adecuado es necesario descubrir qué variables son las que más peso tienen para elaborar el perfil de cliente así como adecuar la dinámica entre dichas variables y los *timing* que se producen en las acciones de marketing lanzadas. La evolución transversal desde el inicio hasta el final se traduce en lo que se conoce como embudo de conversión, que se representa gráficamente con un *funnel chart* dividido entre las diferentes fases por las que evoluciona un cliente potencial.

Repasemos las fases clásicas de un proyecto de implantación *Business Intelligence* para ver, a continuación, en qué fases se centrará el presente TFG.

1. **Definición de objetivos**

   Consiste en analizar las bondades derivadas de este tipo de proyectos y cuáles son realmente las ventajas aplicadas de forma directa a la empresa y su modelo de negocio. Hay que tener en cuenta que estos beneficios pueden ser tangibles y medibles, como una reducción de costes, pero otros quizás sean estratégicos, como aumentar el valor de la marca en el mercado.

   Previamente requerirá de un estudio de situación de la organización desde todas las perspectivas, especialmente evaluando las herramientas actuales de gestión y qué impacto tendrá en ellas la implantación de las nuevas herramientas BI.

   Nuestro proyecto planteará unos objetivos bien definidos. Sin embargo, no realizará un estudio pormenorizado de cada una de las herramientas disponibles en el mercado puesto que no es el principal objetivo del proyecto.

2. **Modelización**

   La información de la fase anterior sienta las bases para modelizar el *Data Warehouse* (almacén de datos) que consiste en el diseño de una base de datos, normalmente bajo el paradigma del álgebra relacional, que dé consistencia a toda la información que maneja la organización así como a otras bases de datos que pudieran ser importadas para enriquecer lo que posteriormente se convertirá en conocimiento. En este modelo de datos se detallaran los metadatos que permitirán en la fase siguiente una extracción eficiente de la información así como el desarrollo de interfaces de conexión para los futuros procesos ETL.

   Esta es una fase crítica de este tipo de proyectos, porque las decisiones pueden influir dramáticamente en sentido positivo o negativo en fases posteriores. El diseño del modelo lógico se traducirá en el modelo físico. En la modelización ya es necesaria la selección y adopción de diferentes herramientas que tendrán un impacto tanto el coste actual como futuro de muchas de las decisiones que se tomen. Se construye, en definitiva, la arquitectura de datos, las fuentes de entrada de información y los

procedimientos para consolidar los datos de forma que su análisis futuro resulte óptimo en términos de velocidad y escalabilidad.

3. Implementación

Una vez establecido el modelo de datos, es hora de realizar varios procesos ETL (extracción, transformación y carga) que alimenten el almacén de datos. Normalmente existe una carga inicial y posteriormente se establecen una serie de procesos periódicos que recaban información de los diferentes canales para seguir agregando datos de forma continua.

Con el almacén de datos operativo, se despliega una maquinaria completa de utilidades para extraer la información y de ésta, el conocimiento. Independientemente del fabricante del software analítico, lo normal es que se haga uso de las siguientes herramientas:

- Utilidades para minería de datos (*Data Mining*).
- Segmentación de datos (Q*uery*) e informes (Reporting).
- Módulos OLAP (On-Line Analytical Processing).
- Sistemas de soporte a decisiones (DSS).
- Cuadros de mando integral (KPI).

En este fase también se incluye un intenso y amplio programa de formación de todos los actores implicados en el proyecto así como el establecimiento de un equipo de soporte técnico para detectar, evitar y corregir posibles fallas en los sistemas implantados.

4. Reevaluación

Como se comentó al inicio de esta introducción, la inteligencia de negocio no es un proceso estático, sino todo lo contrario, requiere de constantes iteraciones para incorporar nuevos datos, ponderar datos ya existentes, revisar validaciones de datos, evaluar la experiencia de usuario y el nivel de adopción de las herramientas en el equipo BI, etc. En definitiva, se busca validar si se han cumplido los objetivos y perfeccionar el método orientando los cambios hacia nuevos objetivos en la estrategia de la organización.

1.2 Justificación del trabajo

Ante la enorme amplitud de fases que implica el desarrollo de un proyecto BI, desarrollar este TFG implica inevitablemente acotar los esfuerzos en algunas de las fases y, además, seleccionar uno de los múltiples modelos de negocio que nos encontramos en el mercado como banco de pruebas de los objetivos que posteriormente se definirán para este trabajo.

Desde mi valoración personal por la experiencia profesional en el desarrollo de este tipo de proyectos, creo que uno de los aspectos de los que adolecen es la falta de tiempo dedicado a la fase de comprensión del modelo de negocio y el diseño de un almacén de datos sólido y escalable. Si esta fase se realiza correctamente, se sentarán los cimientos para cualquier construcción posterior de módulo KPI, OLAP u otro tipo de herramientas gráfica.

Por otro lado, se tiende a elegir una herramienta sin entender bien cuáles son los datos que definen al cliente y, a su vez, cuáles son las diferentes fases que supera un cliente potencial hasta llegar a convertirse en un cliente final satisfecho. Además, entender qué acciones se han realizado en el plan de marketing y su impacto medible en dicha conversión no son fáciles de detectar y cuantificar. En muchas ocasiones se usan técnicas estadísticas muy básicas y gráficos comunes cuando actualmente contamos con numerosos análisis numéricos que no son excesivamente complejos, pero que permiten extraer una información mucho más veraz y acertada para descubrir el peso que tienen las campañas y acciones en la conversión de clientes potenciales.

Un factor de mercado que añade complejidad a elaborar un proyecto BI es el hecho de toparse con un sector dinámico donde el cliente potencial no hace uso de los canales clásicos y su comportamiento muta constantemente. Es el caso de cualquier sector donde el cliente potencial son jóvenes que interactúan con redes sociales y su reclamo no viene de anuncios de TV sino del impacto de *influencers*, aplicaciones de mensajería y canales de vídeos como Youtube. Es en este supuesto donde identificar el perfil medio y cuantificar el ROI de ciertas acciones cobra cierta dificultad y exige un modelo de datos dinámico así como técnicas de análisis más avanzadas.

1.3 Objetivos

El presente trabajo tiene como objetivo principal explorar de forma teórico-práctica cómo sería el desarrollo de un proyecto de BI enfocando principalmente los esfuerzos en el modelado de datos resaltando ciertas decisiones particulares que pueden marcar la diferencia entre el éxito y el fracaso futuro a la hora de plasmar los datos en un KPI.

Para ello se elegirá un sector de mercado dinámico como es el de la educación superior privada. Nos pondremos en la piel de un departamento de marketing cuya función es captar el mayor número posible de alumnos que pueden cursar un grado universitario y analizaremos qué variables definen al cliente potencial, qué variables definen de forma abstracta las diferentes acciones de marketing teniendo en cuenta su secuencia y, sobre todo, cuáles son las fases que sigue el alumno potencial desde su primer contacto hasta la realización de la matrícula.

Una vez consolidado este modelo, se establecerán los principales indicadores así como las técnicas analíticas para obtenerlos que aporten información en la toma de decisiones para lanzar determinadas campañas que ayuden a la consecución de los objetivos de venta.

Paralelamente a esta fase se desarrollará de forma práctica, con una implementación ad-hoc para el modelo de datos, la representación gráfica de algunos recursos innovadores que no suelen usarse de forma habitual pero que constituyen una poderosa herramientas visual. En este sentido nos centraremos en explorar diagramas de flujo como *Sankey Diagram* o gráficos para el progreso del proceso de venta como *Funnel Chart*. Se presentará la base teórica de dichas herramientas, la adecuación al modelo de negocio elegido y varios ejemplos haciendo uso de la base de datos sobre la que se apoyará el presente trabajo.

Finalmente se concluirá con una evaluación del logro de estos objetivos así como de un análisis sobre la líneas futuras de actuación que podrían hacer evolucionar el proyecto a un entorno empresarial más complejo.

1.4 Alcance

Aunque el desarrollo final de este proyecto concluye con un entregable donde se puede valorar la utilidad práctica del modelo de datos y su implementación, dada la naturaleza teórico-práctica de este trabajo, el objetivo es centrarse más en el razonamiento de las decisiones que se toman en cada fase así como en las líneas futuras de actuación teniendo un buen modelo de partida. Es por esto que el alcance del actual trabajo fin de grado serán los siguientes elementos:

- Estudio profundo y razonado del modelo de datos adecuado para servir de base en un proyecto BI orientado a la captación de alumnos para la educación superior.

- Base de datos con información simulada pero cumpliendo con las validaciones y reglas establecidas en el modelo. Para ello se hará uso de procesos ETL.

- Gráficos avanzados que ayuden a medir la consecución de los objetivos de la estrategia de marketing planteada. Se implementarán en el software Power BI conectado a la base de datos del proyecto.

- Líneas de extensión del modelo de datos aplicadas al sector educativo.

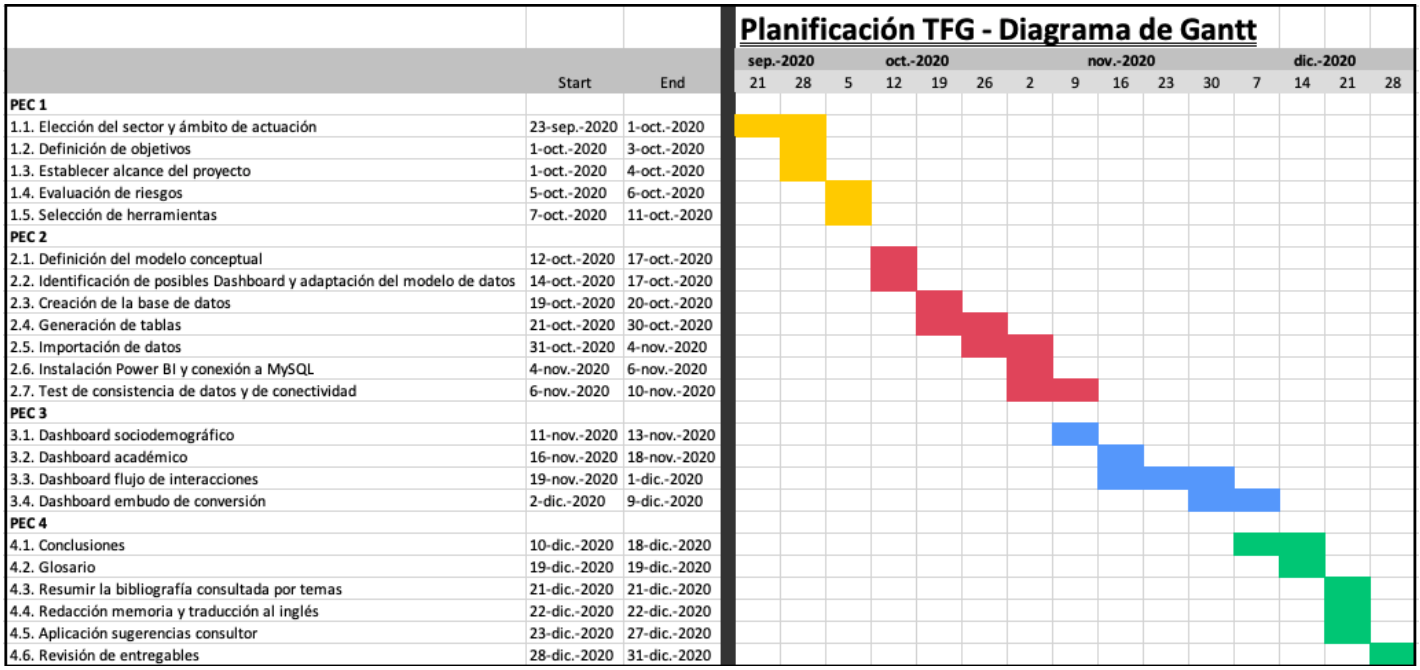
- Conclusiones finales sobre el desarrollo del proyecto y revisión del cumplimiento de los objetivos iniciales.

1.5 Planificación

Se adjunta la planificación temporal de las tareas agrupadas por cada entrega exigida así como desglosada por subtareas cuando ha sido desagregada. Se incluye la prioridad que se estableció inicialmente y que se ha conservado casi intacta a medida que se han ido cerrando tareas.

Planificación TFG					
PEC 1					
Tarea	**Subtarea**	**Estado**	**Comienzo**	**Fin**	**Prioridad**
1.1. Elección del sector y ámbito de actuación	Búsqueda de recursos disponibles como fuentes de datos, Evaluación del potencial analítico de los posibles sectores, Videoconferencia de consultoría	Done	23-sep.-2020	1-oct.-2020	High
	Búsqueda de recursos disponibles como fuentes de datos	Done			
	Evaluación del potencial analítico de los posibles sectores	Done			
	Videoconferencia de consultoría	Done	1-oct.-2020		
1.2. Definición de objetivos		Done	1-oct.-2020	3-oct.-2020	High
1.3. Establecer alcance del proyecto		Done	1-oct.-2020	4-oct.-2020	High
1.4. Evaluación de riesgos		Done	5-oct.-2020	6-oct.-2020	Medium
1.5. Selección de herramientas		Done	7-oct.-2020	11-oct.-2020	Critical
			23-sep.-2020	**11-oct.-2020**	
PEC 2					
Tarea	**Subtarea**	**Estado**	**Comienzo**	**Fin**	**Prioridad**
2.1. Definición del modelo conceptual		Done	12-oct.-2020	17-oct.-2020	High
2.2. Identificación de posibles Dashboard y adaptación		Done	14-oct.-2020	17-oct.-2020	High
2.3. Creación de la base de datos		Done	19-oct.-2020	20-oct.-2020	Medium
2.4. Generación de tablas	Maestros independientes, Programación script tablas dependientes	Done	21-oct.-2020	30-oct.-2020	Critical
	Maestros independientes	Done	21-oct.-2020		
	Programación script tablas dependientes	Done	27-oct.-2020		
2.5. Importación de datos		Done	31-oct.-2020	4-nov.-2020	Critical
2.6. Instalación Power BI y conexión a MySQL		Done	4-nov.-2020	6-nov.-2020	High
2.7. Test de consistencia de datos y de conectividad		Done	6-nov.-2020	10-nov.-2020	Medium
			12-oct.-2020	**10-nov.-2020**	
PEC 3					
Tarea	**Subtarea**	**Estado**	**Comienzo**	**Fin**	**Prioridad**
3.1. Dashboard sociodemográfico		Done	11-nov.-2020	13-nov.-2020	Medium
3.2. Dashboard académico		Done	16-nov.-2020	18-nov.-2020	Medium
3.3. Dashboard flujo de interacciones	Cambios en tabla de interacciones, Reproceso de script para calcular nueva columna 'orden'	Done	19-nov.-2020	1-dic.-2020	Critical
	Cambios en tabla de interacciones	Done			
	Reproceso de script para calcular nueva columna 'orden'	Done			
3.4. Dashboard embudo de conversión		Done	2-dic.-2020	9-dic.-2020	High
			11-nov.-2020	**9-dic.-2020**	
PEC 4					
Tarea	**Subtarea**	**Estado**	**Comienzo**	**Fin**	**Prioridad**
4.1. Conclusiones		Done	10-dic.-2020	18-dic.-2020	High
4.2. Glosario		Done	19-dic.-2020	19-dic.-2020	Low
4.3. Resumir la bibliografía consultada por temas		Done	21-dic.-2020	21-dic.-2020	Low
4.4. Redacción memoria y traducción al inglés		Done	22-dic.-2020	22-dic.-2020	Critical
4.5. Aplicación sugerencias consultor		Done	23-dic.-2020	27-dic.-2020	High
4.6. Revisión de entregables		Done	28-dic.-2020	31-dic.-2020	High
			10-dic.-2020	**31-dic.-2020**	

A continuación, se muestra dicha planificación como diagrama de Gantt:



	Start	End
PEC 1		
1.1. Elección del sector y ámbito de actuación	23-sep.-2020	1-oct.-2020
1.2. Definición de objetivos	1-oct.-2020	3-oct.-2020
1.3. Establecer alcance del proyecto	1-oct.-2020	4-oct.-2020
1.4. Evaluación de riesgos	5-oct.-2020	6-oct.-2020
1.5. Selección de herramientas	7-oct.-2020	11-oct.-2020
PEC 2		
2.1. Definición del modelo conceptual	12-oct.-2020	17-oct.-2020
2.2. Identificación de posibles Dashboard y adaptación del modelo de datos	14-oct.-2020	17-oct.-2020
2.3. Creación de la base de datos	19-oct.-2020	20-oct.-2020
2.4. Generación de tablas	21-oct.-2020	30-oct.-2020
2.5. Importación de datos	31-oct.-2020	4-nov.-2020
2.6. Instalación Power BI y conexión a MySQL	4-nov.-2020	6-nov.-2020
2.7. Test de consistencia de datos y de conectividad	6-nov.-2020	10-nov.-2020
PEC 3		
3.1. Dashboard sociodemográfico	11-nov.-2020	13-nov.-2020
3.2. Dashboard académico	16-nov.-2020	18-nov.-2020
3.3. Dashboard flujo de interacciones	19-nov.-2020	1-dic.-2020
3.4. Dashboard embudo de conversión	2-dic.-2020	9-dic.-2020
PEC 4		
4.1. Conclusiones	10-dic.-2020	18-dic.-2020
4.2. Glosario	19-dic.-2020	19-dic.-2020
4.3. Resumir la bibliografía consultada por temas	21-dic.-2020	21-dic.-2020
4.4. Redacción memoria y traducción al inglés	22-dic.-2020	22-dic.-2020
4.5. Aplicación sugerencias consultor	23-dic.-2020	27-dic.-2020
4.6. Revisión de entregables	28-dic.-2020	31-dic.-2020

Finalmente, se adjunta captura de la herramienta online “monday” que ha sido utilizada tanto para la elaboración de la planificación, como para el seguimiento de las tareas con la metodología kanban.

1.6 Riesgos

Dadas las aspiraciones que se expresan en los objetivos, cabe analizar cuanto menos los riesgos asociados teniendo en cuenta el *timing* de la planificación. Se han detectado los siguientes riesgos:

Riesgo	Incumplimiento de timing.
Causa	Desconocimiento previo de la herramienta Power BI tanto en lo referente a la interconexión con fuentes de datos como al uso de DAX y creación de gráficos avanzados.
Efecto	Obligación a reducir el alcance y los objetivos.
Probabilidad	Media.
Acción a realizar	Familiarizarse con Power BI probando con fuentes de datos alternativas existentes para crear gráficos similares. Si se percibe que no es la herramienta adecuada, probar otros software de mercado.

Riesgo	Incumplimiento de objetivos iniciales.
Causa	Bien por un error en el modelo de datos o por una carencia de recursos en Power BI. El modelo de datos requiere de un diseño muy concreto para poder elaborar los análisis que se han planteado, pero hasta no llevarlo a cabo, hay cierta probabilidad de que no sea compatible con los objetivos. Especialmente nos referimos al Dashboard de interacciones y el funnel de conversión.
Efecto	Pérdida de tiempo y afectación en la planificación al tener que realizar cambios en ETL y/o alcanzar un resultado distinto al planteado en los objetivos.
Probabilidad	Alta
Acción a realizar	Familiarizarse con Power BI probando con fuentes de datos alternativas existentes para crear gráficos similares.

Riesgo	Redacción de difícil lectura.
Causa	El área de marketing hace un uso intensivo de anglicismos y acrónimos.
Efecto	Dificultad de comprensión tanto de los objetivos planteados como de las conclusiones.
Probabilidad	Baja
Acción a realizar	Minimizar el uso de anglicismos y aclarar, tanto en el texto principal como en el glosario aquellos términos propios del modelo de negocio para facilitar la lectura del proyecto.

Riesgo	Cuadros de mando difíciles de utilizar e interpretar
Causa	Excesiva complejidad no tanto en los objetivos planteados sino en el desarrollo y resultado final de estos cuadros de mando.
Efecto	Convertir los resultados de la implementación en herramientas innecesarias, inútiles o incomprensibles para el rol del director de marketing.
Probabilidad	Baja
Acción a realizar	Alcanzar un compromiso entre la simplificación del cuadro de mando manteniendo cierto grado de control con algunos filtros pero evitando la sobrecarga de selectores y multitud de gráficos que pudieran ser redundantes o innecesarios.

Capítulo 2 – Definiendo la base conceptual

2.1 Entendiendo el modelo de negocio

El departamento de marketing de un centro educativo debe entender bien cómo es su perfil de alumno potencial. Para ello será necesario averiguar qué variables definen a este perfil. Aunque posteriormente se profundizará en este aspecto, vemos *grosso modo* que un alumno es una persona que viene determinada por características sociales (edad, género, nivel educativo previo) y económicas, que a su vez vienen condicionadas por aspectos como su ubicación, el nivel de renta de su familia, etc.

Una vez comprendido este perfil, la labor de marketing pasa por administrar una serie de recursos disponibles en realizar labores de captación por diferentes canales. Estas labores se traducen de forma efectiva en acciones concretas activas como pueden ser el envío de un SMS con redirección a una *landing page,* o la convocatoria en redes sociales a un evento presencial como, por ejemplo, una jornada de puertas abiertas. También acciones pasivas como la recepción de llamadas telefónicas para aportar información o asesorar a los alumnos potenciales.

A su vez, será necesario dividir las fases del proceso de venta de forma que se adecúe al modelo concreto de negocio. Normalmente, y de forma genérica, se establecen las siguientes fases:

- Adquisición
- Activación
- Retención
- Venta
- Referencia

Estas fases están muy vinculadas al comercio electrónico. Por lo tanto, en el sector concreto que nos ocupa, vamos a adaptar dichas fases de forma que se encajen en el modelo educativo de la siguiente forma:

- **Interesado.** Solicita información por algún canal.
- **Activado.** Demuestra mayor nivel de interés.
- **Evento presencial.** Se inscribe para visitar el campus o asistir a algún otro evento como feria educativa.
- **Admisión.** Realiza el proceso de admisión.
- **Matrícula.** Supera la admisión, formaliza una matrícula y realiza su pago.

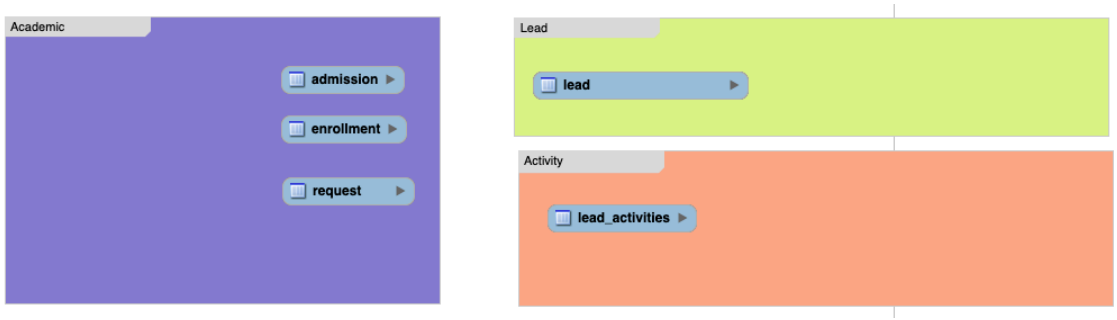
Obviamente, detrás de cada fase (especialmente de la primera), hay grandes volúmenes de alumnos cada uno con su propia interacción con las determinadas acciones realizadas. Normalmente los sistemas CRM de una organización aplican procesos de normalización y deduplicación de datos con el objetivo de identificar de forma unívoca al cliente potencial. Este paso es imprescindible para poder realizar una trazabilidad de cada caso en el embudo de ventas y tener esa visión 360 típica de una buen sistema de gestión que apoye al sistema BI analítico. En el modelo que se desarrollará a continuación se dará por hecho que existe esta identificación.

Otro aspecto importante a tener en cuenta es qué información solicitan los alumnos potenciales. Podrían estar interesados por multitud de aspectos como residencias, becas, descuentos, etc. Sin embargo, el principal dato que obtendremos de sus peticiones de información será la titulación que desean cursar. Esta titulación pertenecerá a un área concreta, ya que las acciones de marketing emprendidas podrán ser distintas para alumnos que quieran cursar grados de la rama técnica a aquellos que deseen estudiar carreras de humanidades. Sus perfiles socioeconómicos, intereses y aficiones podrían ser muy distintos y tener ese dato permitirá afinar más las campañas, por no decir que cada área podría tener expertos en marketing distintos que opten por realizar campañas propias segmentando a los potenciales de sus ramas.

En esta narrativa se empiezan a descubrir con cierta facilidad cuáles son las entidades que van a intervenir en el modelo relacional que consolidará la información del proyecto:

- Información del alumno potencial
- Interacciones con el alumno potencial
- Peticiones de información, admisiones y matriculaciones

Antes de pasar a definir en detalle cada una de ellas, veamos un boceto previo de estas tres partes en el siguiente esquema.



- Vemos claramente en la sección “Lead” cómo existirá una entidad Lead que almacenará los datos de los potenciales.
- La actividad del centro con los leads se registrará en la entidad lead_activities.
- La actividad académica de interés se registrará en 3 entidades: request (peticiones), admission (admisiones), enrollment (matrículas).

A continuación, en la definición formal de cada una de ellas, veremos cómo será necesario definir otras entidades para definir ciertos atributos de las entidades principales del modelo. También será imprescindible relacionar cada una de ellas.

En este proceso se aplicarán las formas normales, aunque intentando conseguir un compromiso entre los siguientes aspectos:

- Normalización de los datos.
- Escalabilidad del modelo.
- Eficiencia para las obtención de datos desde la perspectiva del lenguaje SQL.

Pasemos, por tanto, a definir formalmente cada una de ellas en el siguiente apartado.

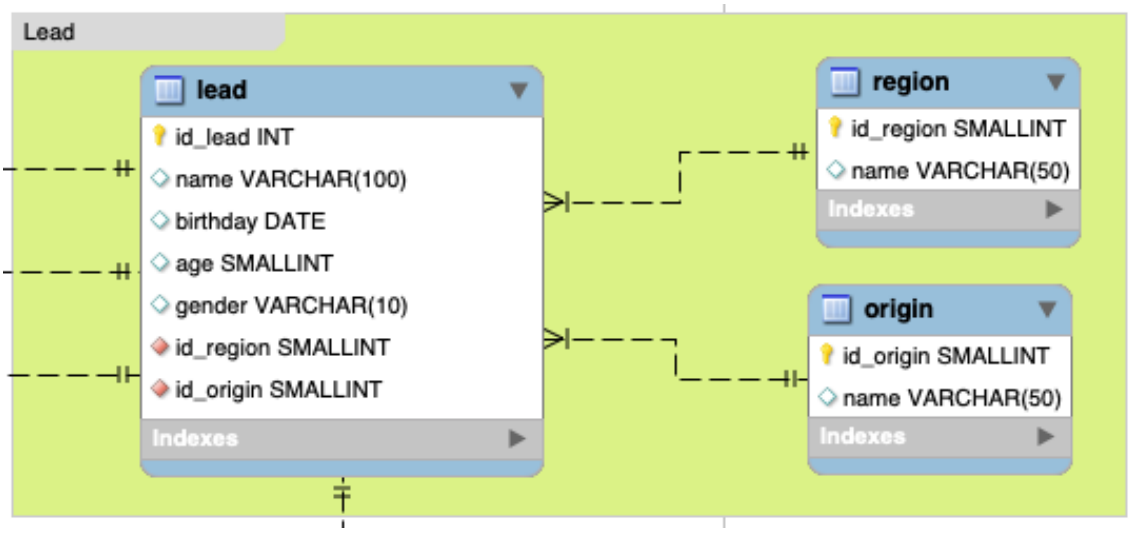
2.2 Definición de las entidades

2.2.1 El Alumno potencial (*lead*)

Por usar terminología de marketing, llamaremos *lead* al alumno potencial. Se entiende por *lead* al usuario que por algún canal ha cedido sus datos, en este caso, al centro educativo.

Se simplificará bastante la entidad aunque, como se comentó en la introducción, se resaltarán algunas decisiones de diseño que afectarán de forma positiva y negativa al conocimiento que podremos derivar en fases posteriores.

La entidad lead la definimos de esta forma:



Los campos son los siguientes:

Campo	**Apreciaciones**
id_lead	Identificador
name	Nombre del lead
birthday	Fecha de nacimiento
age	Edad
gender	Género
id_region	Código de provincia
id_origin	Procedencia curricular del lead

Vemos en la entidad algunas decisiones tomadas con intención de evidenciar la influencia que estas tendrán en extraer la información o almacenar nuevos datos.

- Campo *birthday*

  Si almacenamos la fecha de nacimiento, podremos recalcular la edad del lead con exactitud a medida que pasan los días. Si pidiéramos sólo la edad en los formularios de solicitud de información, este dato quedaría obsoleto y no podríamos hacer uso del lead en el tiempo o bien recalcularlo sólo de forma aproximada.

- Campo *age*

  Este campo también tiene su particularidad a la hora de incluirlo o no. Si lo incluimos, tenemos la ventaja de que podemos realizar consultas sobre él sin tener que aplicar fórmulas en tiempo real sobre la fecha de nacimiento. Si no sabemos todavía el software analítico que usaremos y desconocemos si se podrían aplicar esas funciones, es mejor tenerlo ya calculado. Por el contrario, de contar con él, necesitaremos un proceso periódico diario para recalcular este campo en el almacén de datos.

- Campo *gender*

  Vemos que se trata de una cadena de texto. No vamos a enlazar este campo con ningún maestro. La ventaja la tenemos de cara a unas consulta más veloces. La desventaja es que no podremos incorporar nuevos géneros ya que restringiremos los posibles valores a dos. Entendemos que no será algo necesario y dado el número tan limitado y estático de valores, prescindiremos de un maestro.

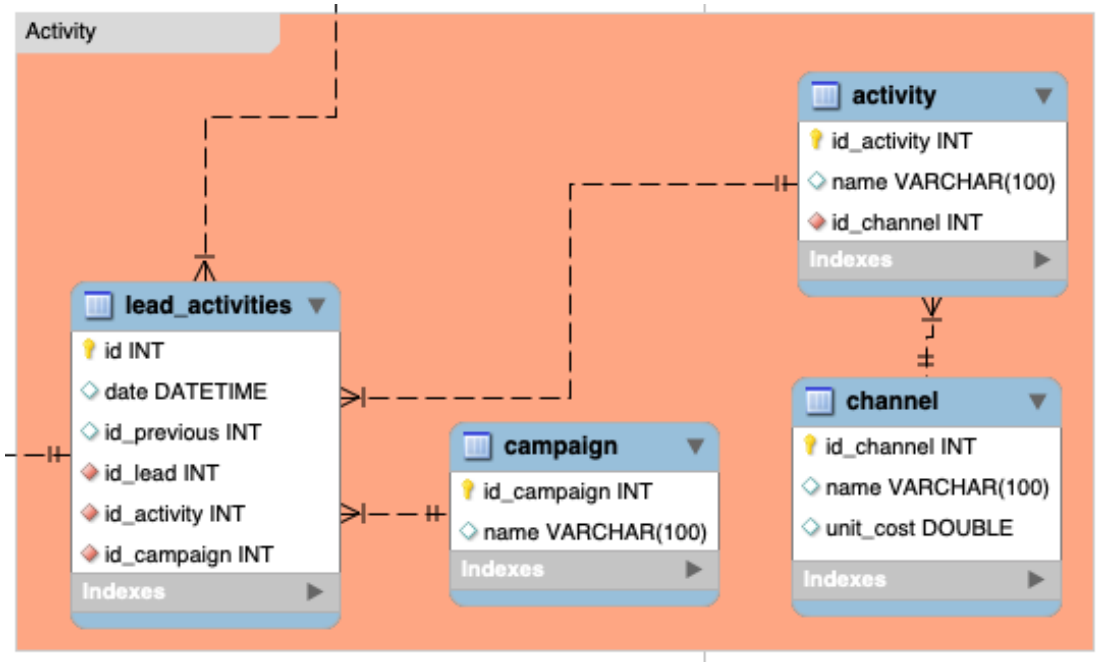
- Campo *id_region*

  Al igual que con el campo "id_origin" que aparece el último de la entidad, aquí sí vamos a introducir una clave foránea con el maestro de regiones que es donde se almacenarán las provincias. No vamos a entrar en construir una estructura más compleja multi-país ya que el objetivo del proyecto no se enfoca en su totalidad en el modelo de datos. Pero sí vamos a variar esta decisión con respecto al campo del género para ver cómo influyen estas decisiones en la parte BI posterior que extrae información.

2.2.2 La actividad del alumno en la fase de captación

Aquí podríamos complicar el modelo, pero vamos a alcanzar un compromiso entre el mundo real y las necesidades del proyecto de cara a conseguir gráficos potentes que ubicar en un KPI. Por esta razón, vamos a considerar que el departamento de marketing cuenta con una serie de canales para comunicarse e interactuar con el alumno potencial y viceversa. Y, basándonos en esos canales, existirán diferentes acciones a realizar. Un siguiente nivel superior sería asociar determinadas acciones concretas en una campaña y así poder medir la efectividad de las campañas de forma separada. Veremos más tarde cómo este nivel superior puede ser útil para nuestros objetivos.

Un dato muy interesante que vamos a incluir en el esquema de interacciones es la acción o actividad precedente a una acción. Este dato puede ser muy revelador de cara a los diagramas de flujo que nos den pistas sobre qué interacciones provocan otras interacciones. Por ejemplo, podemos tener una campaña en redes que derive en una *landing page* y en esta *landing page* una inscripción en un evento presencial. O bien lanzar un *emailing* a cierto segmento de la base de datos de potenciales donde les dirijamos a un blog. O un lanzamiento de una campaña de SMS con un número de teléfono al que llamar. Las combinaciones son infinitas. Pero si podemos trazar el vínculo entre unas y otras, seremos capaces de visualizar el flujo. El modelo propuesto es el siguiente:



Contamos con un maestro de canales (*channel*), maestro de campañas (*campaign*) y maestro de actividades (*activity*). Finalmente tenemos la tabla que registra cada interacción con el lead (*lead_activities*). Veamos algunas particularidades del esquema.

- Campo *lead_activities.id_previous*

  En este campo vinculamos, como si de una lista enlazada se tratara, la actividad con su actividad anterior. No hemos formado una clave foránea debido a que es un dato que no siempre podremos tener.

- Campo *lead_activities.id_campaign*

  Enlazamos la actividad con una campaña. Quizás haya actividades huérfanas pero para ello usaremos campañas genéricas con tal de poder agruparlas todas. Por ejemplo, tendremos una campaña muy concreta como es el lanzamiento de una *landing page* que invite a un evento presencial y ahí sí podremos englobarlas claramente. Pero si un interesado se pone en contacto con el *contact center* de forma espontánea, agruparemos esa actividad a una campaña genérica de atención al cliente.

- Campo *activity.id_channel*

  En el maestro de tipos de actividades, vemos que cada tipo tiene como propiedad el canal por el que se realiza. No puede haber ninguna actividad que no pertenezca a ningún canal. Todas deben tener ese campo que nos permitirá, como veremos en el siguiente campo que analizamos, calcular los costes totales.

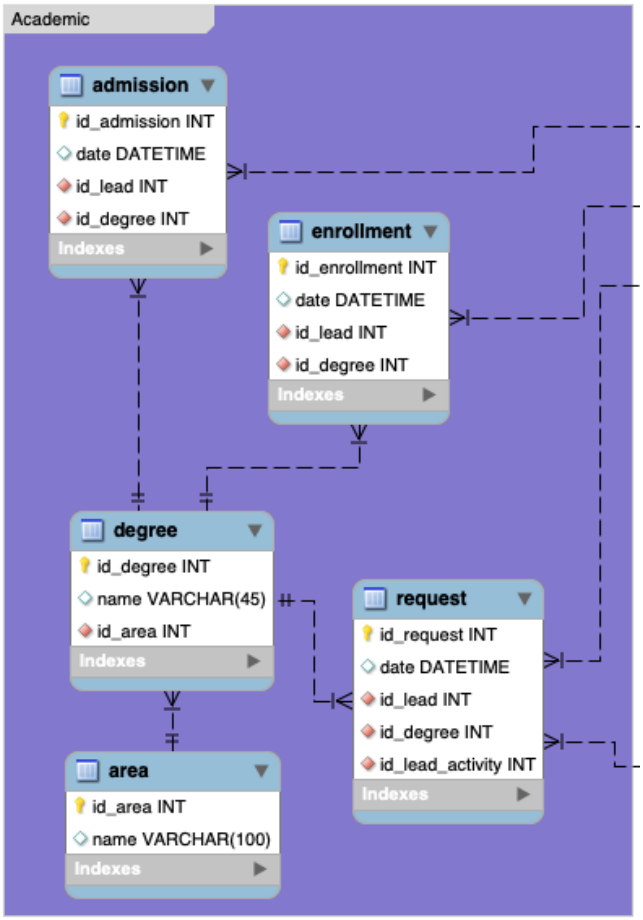
- Campo *channel.unit_cost*

  Aunque de forma aproximada, aquí podremos indicar el coste unitario de cada actividad. Hay casos en los que el coste será exacto, como el envío de un SMS. En otras ocasiones el departamento de marketing indicará una estimación como puede ser el coste de atención de llamada telefónica o la visita de un alumno potencial a un evento.

2.2.3 Interés del alumno y avance en el proceso de venta

En las primeras interacciones que el alumno potencial tenga con el centro educativo, lo importante es recoger qué titulación es la de interés. Podrían ser varias pero vamos a simplificar el modelo recogiendo sólo una. Hay varias opciones en el modelo relacional para registrar esta información. Una opción sería en la actividad o interacción concreta donde se produce. Pero quizás sea mejor tener una entidad donde recoger este dato aunque podamos vincularlo a la actividad concreta.

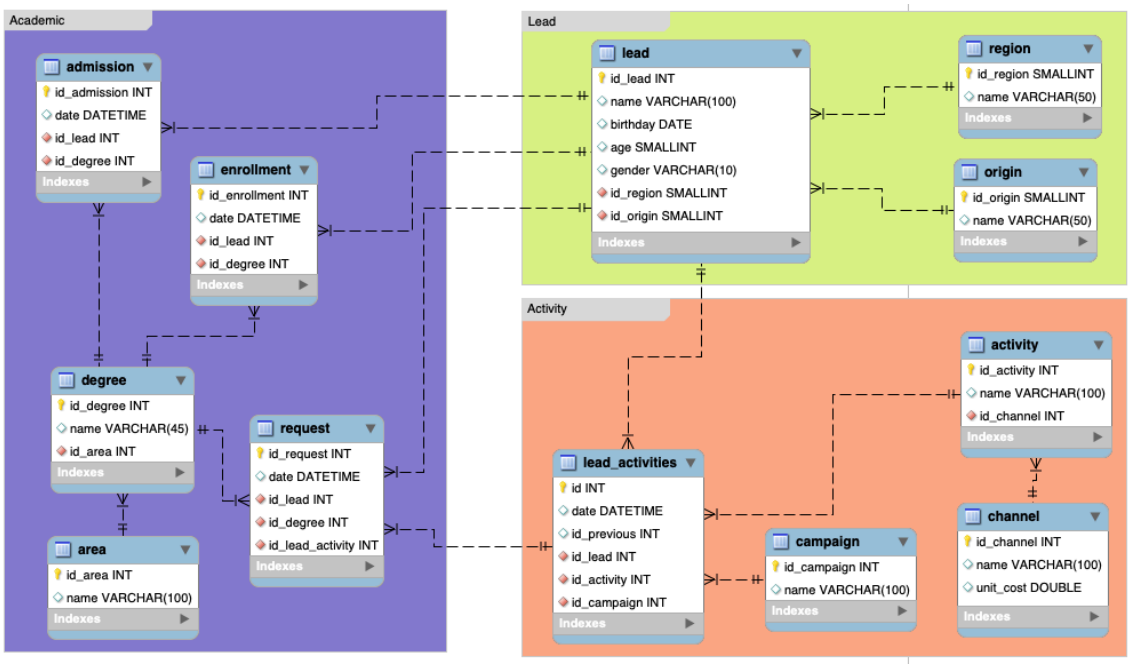
Por otro lado, también será necesario registrar cuándo se produce la admisión y la matriculación si es que el lead avanza en el proceso de venta hasta ese punto. Aquí también la fecha es importante de cara a análisis posteriores. Veamos cómo queda esta parte del modelo que incluye el maestro de titulaciones, donde tendremos cada una agrupada por áreas.



Se ha incluido alguna redundancia en esta parte del modelo cuando en la tabla de peticiones de información (*request*) se registra el identificador del lead que se podría haber obtenido gracias al identificador de la actividad del lead. Sin embargo, cuando queramos analizar el embudo de ventas sin tener en cuenta las interacciones, será más eficiente la consulta de esta forma. Igualmente, y sólo para las peticiones de información, se enlaza con la interacción que ha provocado dicha petición. No se hará lo mismo con la admisión (*admission*) y la matriculación (*enrollment*) ya que no es relevante desde el punto de vista de la captación (si bien existen acciones propias para el hecho de realizar admisión y matrícula).

En las tres tablas se registra la titulación solicitada cuyo maestro se encuentra en la tabla *degree* y, como comentamos antes, varias se agrupan en un área (tabla *area*) concreta.

El modelo completo queda de la siguiente forma:

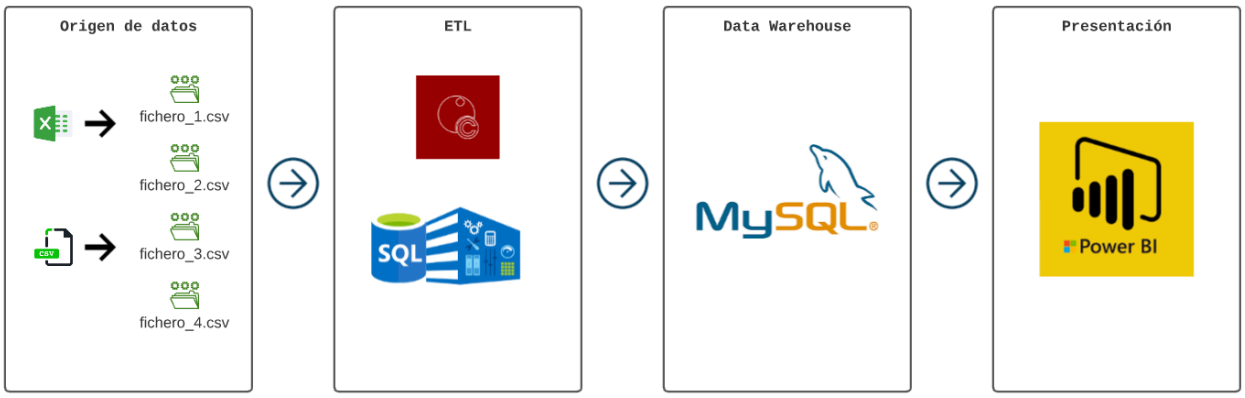


El modelo queda relacionado cumpliendo con los criterios de normalización. Se han separado en tres áreas conceptuales que ayudan a visualizar cada sección del proyecto.

Capítulo 3 – Construcción del Data Warehouse

3.1 Arquitectura del equipo y soluciones tecnológicas adoptadas

El esquema completo por fases de las diferentes herramientas utilizadas es el siguiente:



A continuación se enumera cada elemento de la arquitectura. Comencemos viendo las características del equipo utilizado para implementar el proyecto.

Equipo

Dado que la plataforma Power BI no está disponible en sistemas operativos macOS, se ha optado por el uso de un PC con el sistema operativo Windows 10. Las características concretas son las siguientes:

Procesador	Interl® Core™ i5-3330 CPU 3.00GHz
Memoria RAM	8,00 GB
Arquitectura	64 bits
Versión Windows	10 Pro 18362.1016
Disco duro	SSD 256 GB

Herramientas Redacción y Diseño

- **MarkdownPad**. Se ha utilizado para el borrador en la redacción del texto principal.
- **Microsoft Word 2010**. Para la maquetación del texto y las imágenes.

- **Lucidchart**. Herramienta web que permite elaborar varios tipos de diagramas. Se ha utilizado para diseñar los esquemas gráficos.

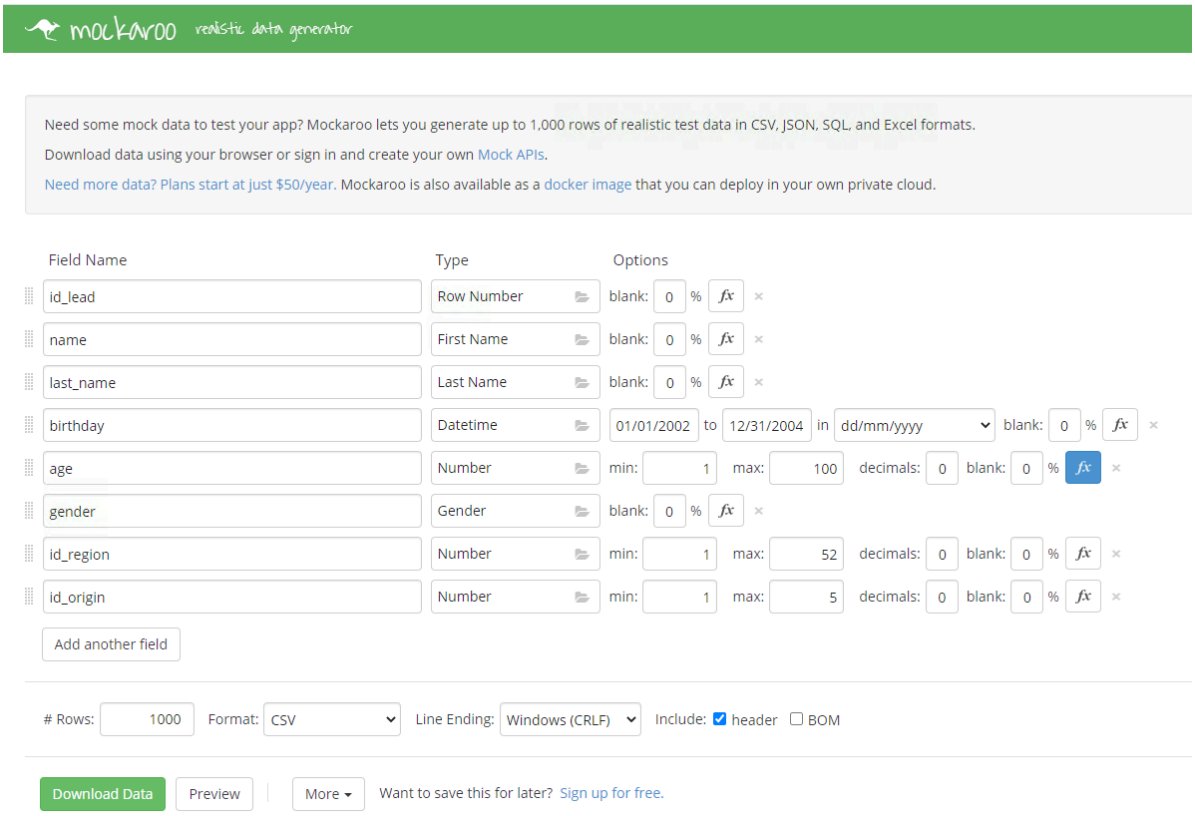
Aplicaciones de modelado y ETL

- **MySQL Workbench 8.0**. Para el diseño del modelo relacional así como para gestionar la base de datos del Data Warehouse se ha hecho uso de esta herramienta visual que integra estas funcionalidades necesarias para trabajar con datos. Está desarrollado por Oracle Corporation y tiene licencia GNU.

- **Microsoft Excel**. Hoja de cálculo. Se ha hecho uso para la generación de los datos simulados en las tablas maestras del modelo relacional y generar los archivos CSV necesarios para importarlos a la base de datos MySQL.

- **Cosmos**. Herramienta RAD (Rapid Application Development) con motor propio de base de datos (Multibase) propiedad de la empresa española Base100. Se ha usado este entorno para la generación de los datos simulados en aquellas tablas no maestras donde había que aplicar un proceso con funciones aleatorias controladas emulando la similitud con datos reales.

- **mockaroo**. Página web que permite la generación de datos y su exportación en formato plano o instrucciones de inserción en base de datos. Se ha utilizado para la generación de datos en la tabla *Lead* del modelo relacional. A continuación se adjunta una imagen del la configuración establecida para la generación de datos.



Data & Metadata Storage

- **MySQL**. Sistema de gestión de base de datos relacionales propiedad de Oracle Corporation, de licencia GPL.

Software de análisis y presentación

- **Power BI**. Servicio de Business Intelligence propiedad de Microsoft que proporciona visualizaciones interactivas junto con potentes herramientas de importación y tratamiento de datos. Permite crear paneles, informes y gráficos.

3.2 Creación de la base de datos

3.2.1 Generación de ficheros

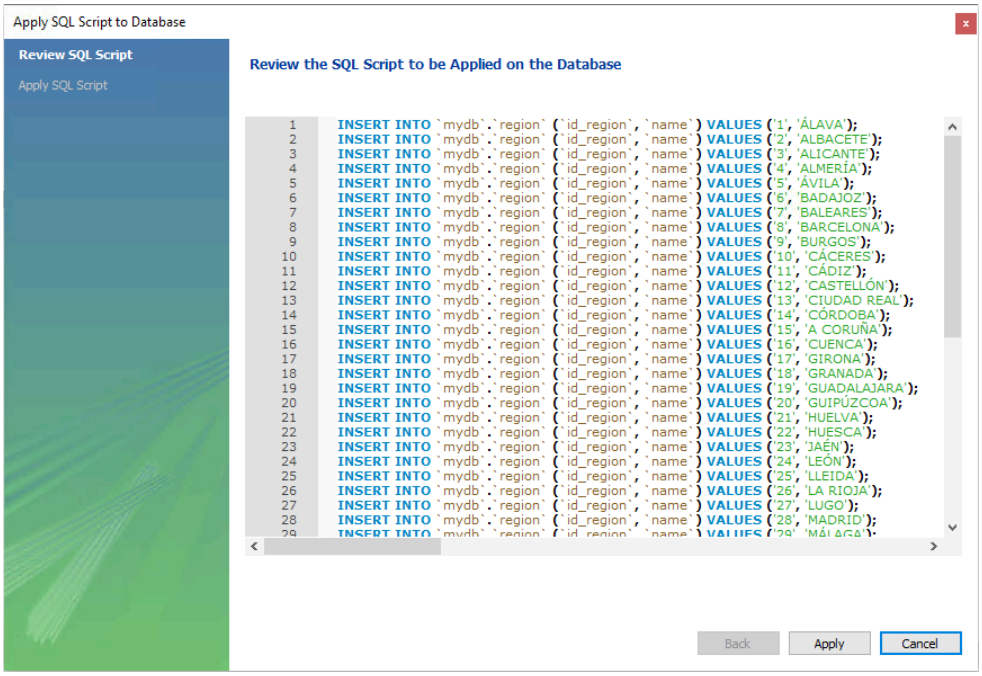
A continuación se detalla el proceso de creación de la base de datos. Se describirá cada tabla así como el orden de importación ya que, debido a las restricciones propias de las claves foráneas, es necesario importar en orden los datos de cada tabla.

Creación de maestros

De las doce tablas que componen el modelo de datos, hay siete que son tablas maestras. A continuación se indican los datos que se han considerado para cada una de ellas:

- `region`

Contiene las 50 provincias y 2 ciudades autónomas de España. La información está disponible desde varias fuentes aunque se ha elegido la publicada de forma gratuita en la servicio de códigos postales y direcciones *geopostcodes*. A continuación se muestra una captura de la importación de los datos en MySQL Workbench. Previamente a este paso se importaron los datos y Workbench generó las sentencias SQL de inserción de forma automática.



- *origin*

Contiene el nivel de estudios previos del lead que llamaremos origen. Se han determinado los siguientes orígenes:

id_origin	name
1	BACHILLERATO
2	CICLO DE GRADO SUPERIOR
3	GRADO UNIVERSITARIO
4	MAYORES DE 25
5	MAYORES DE 45

- *channel*

Los canales de comunicación entre el centro educativo y el lead junto con el coste medio unitario en euros de interacción por este canal. Recordamos que cada interacción tiene que pertenecer a un canal. Se establecen los siguientes:

id_channel	Name	unit_cost
1	LANDING PAGE	1.0
2	EMAIL	0.01
3	SMS	0.04
4	LLAMADA ENTRANTE	3.0
5	LLAMADA SALIENTE	2.0
6	EVENTO PRESENCIAL	5.0

- *activity*

Las diferentes actividades o interacciones que pueden darse entre el centro educativo y el lead. Se establecen las siguientes aunque no todas son usadas:

id_activity	name	id_channel
1	Landing bienvenida	1 (LANDING PAGE)
2	Landing autorespondedor	1 (LANDING PAGE)
3	Landing inscripción AULA	1 (LANDING PAGE)
4	Emailing selectividad	2 (EMAIL)
5	Emailing AULA	2 (EMAIL)
6	Emailing pasos matrícula	2 (EMAIL)
7	SMS petición información	3 (SMS)
8	SMS confirmación admisión	3 (SMS)
9	SMS confirmación matrícula	3 (SMS)
10	Llamada petición información	4 (LLAMADA ENTRANTE)
11	Llamada reclamación	4 (LLAMADA ENTRANTE)
12	Llamada autorespondedor	5 (LLAMADA SALIENTE)
13	Llamada comunicación apto	5 (LLAMADA SALIENTE)
14	Feria AULA	6 (EVENTO PRESENCIAL)
15	Jornada de puertas abiertas	6 (EVENTO PRESENCIAL)
16	Realiza admisión	6 (EVENTO PRESENCIAL)
17	Realiza matrícula	6 (EVENTO PRESENCIAL)

- *campaign*

La campaña se corresponde con el año académico donde se agrupa toda la actividad de captación. Por no complicar el modelo no se ha introducido un segundo nivel que podría corresponder a secuencias de acciones de captación. Por ejemplo, englobar todo lo relacionado con la Feria AULA (Landing page, llamadas, evento) en una subcampaña de un año académico. Sólo se establece una campaña para el modelo:

id_campaign	name
1	2020/2021

- *area*

Poniendo en común diferentes áreas similares de varias universidades españolas, se establecen las siguientes (a la derecha el fichero CSV que se carga en Workbench):

id_area	name
1	CIENCIAS SOCIALES
2	CIENCIAS DE LA INFORMACIÓN
3	BIOSANITARIAS
4	ARQUITECTURA
4	INGENIERÍA
5	EDUCACIÓN

```
1,CIENCIAS SOCIALES
2,CIENCIAS DE LA INFORMACIÓN
3,BIOSANITARIAS
4,ARQUITECTURA
5,INGENIERÍA
6,EDUCACIÓN
```

- *degree*

Al igual que con las áreas de estudio, se han incluido para cada área una serie de grados o titulaciones. Se muestran algunas de ellas (son en total 22):

id_degree	name	id_area
1	ADE	1 (CIENCIAS SOCIALES)
2	MARKETING	1 (CIENCIAS SOCIALES)
3	RELACIONES INTERNACIONALES	1 (CIENCIAS SOCIALES)
4	DERECHO	1 (CIENCIAS SOCIALES)
5	BELLAS ARTES	2 (CIENCIAS DE LA INFORMACIÓN)
6	DISEÑO Y PRODUCCIÓN DE VIDEOJUEGOS	2 (CIENCIAS DE LA INFORMACIÓN)
7	COMUNICACIÓN AUDIOVISUAL	2 (CIENCIAS DE LA INFORMACIÓN)
8	PERIODISMO	2 (CIENCIAS DE LA INFORMACIÓN)
9	PUBLICIDAD	2 (CIENCIAS DE LA INFORMACIÓN)
10	BIOTECNOLOGÍA	3 (BIOSANITARIA)
11	ENFERMERÍA	3 (BIOSANITARIA)
12	FARMACIA	3 (BIOSANITARIA)
13	FISIOTERAPIA	3 (BIOSANITARIA)
14	MEDICINA	3 (BIOSANITARIA)
15	PSICOLOGÍA	3 (BIOSANITARIA)
16	ARQUITECTURA	4 (ARQUITECTURA)
17	INGENIERÍA INFORMÁTICA	5 (INGENIERÍA)

Creación de tablas con datos elaborados

- *lead*

Aunque realmente se podría considerar una tabla maestra, dada su estructura así como el papel que juega en el modelo, se define en este apartado. Como se comentó previamente, se ha usado un servicio online para generar datos inventados. Se muestra un ejemplo del fichero CSV compuesto por 1198 leads:

```
1,Jenn Spinola,04/10/2004,16,Female,8,1,
2,Lilla Robertazzi,20/07/2004,16,Female,8,1,
3,Dawna Searl,04/01/2001,19,Female,8,2,
4,Linoel Fellgatt,13/12/2000,19,Male,8,2,
5,Doralynn Grayson,10/01/2001,19,Female,8,2,
```

- *lead_activities, request, admission y enrollment*

Esta es una de las partes más complicadas del ETL dado que necesita establecer ciertos sesgos en la información generada, hacer uso de los maestros así como de aplicar funciones aleatorias para realizar una distribución no uniforme que permita dar sentido a los gráficos mostrados en Power BI.

Se ha escrito un script en Cosmos que recorre los leads y elije, de una colección previa de posibles secuencias de interacciones, una de ellas e inserta en las diferentes tablas la información correspondiente. También aplica una probabilidad de 1 entre 10 de que un lead realice admisión. Si es así, se aplica una probabilidad de 9 entre 10 de que se matricule. Por supuesto, todos los lead tendrán una petición de información.

Además, se ha establecido diferencias en las secuencias de interacción entre aquellos leads provenientes de Bachillerato y Ciclo superior y los de mayor edad provenientes de graduados, mayores de 25 años y mayores de 45 años. De hecho, en la creación en *mockaroo* se realizan varios ficheros para aplicar unos rangos de edad distintos en función del estudio previo de lead. De esta forma, cuando se analicen los datos se podrán observar tendencias en las variables socioeconómicas que definen al lead relacionadas con las académicas.

Se ha elaborado un pseudocódigo del script que se muestra a continuación:

```
script generacionTablasDependientes
        conectar database

        crear colección A (bach y ciclos)
        añadir elemento 1 colección A
        añadir elemento 2 colección A
        añadir elemento 3 colección A
        añadir elemento 4 colección A

        crear colección B (grados y mayores 25,45)
        añadir elemento 1 colección B
        añadir elemento 2 colección B
        añadir elemento 3 colección B
        añadir elemento 4 colección B

        recorrer leads
                asignar tiene_admision (boolean. Cierto si aleatorio entre 1 y 10 == 1) // 10% de probabilidad
                si tiene_admision entonces
                        asignar tiene_matricula (boolean. Cierto si aleatorio entre 1 y 10 <> 1) // 90% de probabilidad
                fin si

                asignar_titulacion (aleatorio entre 1 y 22 titulaciones del maestro)

                asignar_fecha (aleatorio entre enero y marzo 2020)

                evaluar origen
                        caso origen bachillerato y ciclos
                                elegir elemento aleatorio colección A
                                recorrer pasos de la secuencia
                                        insertar actividad (tabla lead_activities)
                                        guardar id actividad
                                        sumar fecha (entre 1 y 15 días aleatorio)
                                        si es primer paso entonces
                                                insertar petición (tabla request)
                                        fin si
                                fin recorrer
                        fin caso
                        caso grados y mayores
                                elegir elemento aleatorio colección B
                                recorrer pasos de la secuencia
                                        insertar actividad (tabla lead_activities)
                                        guardar id actividad
                                        sumar fecha (entre 1 y 15 días aleatorio)
                                        si es primer paso entonces
                                                insertar petición (tabla request)
                                        fin si
                                fin recorrer
                        fin grado
                fin evaluar

                si tiene_admision entonces
                        sumar fecha (entre 1 y 15 días aleatorio)
                        insertar actividad comunicación apto
                        guardar id actividad
                        sumar fecha (entre 1 y 15 días aleatorio)
                        insertar actividad de admisión
                        guardar id actividad
                        insertar admision (tabla admission)
                fin si

                si tiene_matricula entonces
                        sumar fecha (entre 1 y 15 días aleatorio)
                        insertar actividad de matrícula
                        insertar matrícula (tabla enrollment)
                fin si
        fin recorrer leads
        cerrar database
fin script
```

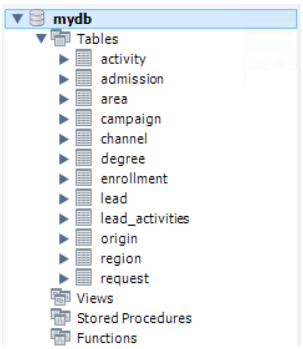
3.2.2 Importación de ficheros y descripción del proceso ETL

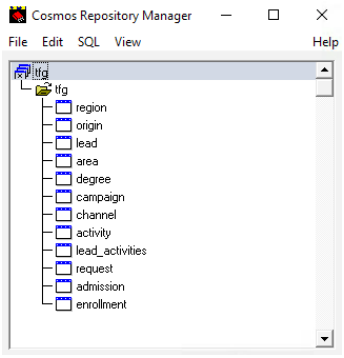
1. El proceso comienza creando desde mockaroo el fichero de leads y exportándolo en formato CSV.
2. A continuación se crean en Excel de forma manual los maestros independientes. La serialización se realiza con herramientas de autocompletado de Excel.
3. Se exportan todos los maestros a formato CSV.
4. Desde Cosmos, se importan a cada tabla los ficheros.
5. En Cosmos se ejecuta el script que genera las tablas *lead_activities, request, admission y enrollment*.
6. Con todas las tablas con datos correctos y serializados, se exportan las cuatro tablas generadas en el proceso.
7. Finalmente en MySQL se ejecuta el script de importación para cada tabla en el orden correspondiente para que se cumplan las reglas de integridad referencial.

A continuación se indica el código de ejemplo de una parte del script de importación donde se realiza una transformación del formato de fecha ya que en Cosmos es distinto al formato de MySQL. Los ficheros CSV se ubican dentro de la carpeta de la base de datos en MySQL para que sea accesible con la siguiente instrucción:

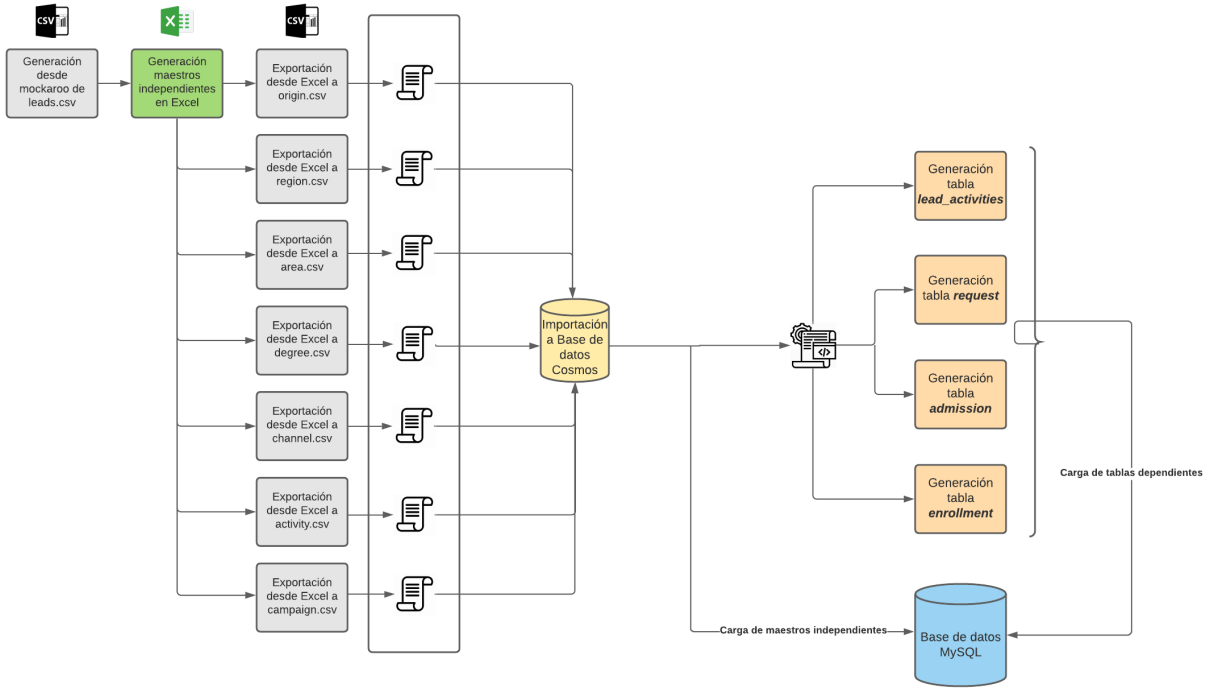
```
LOAD DATA INFILE 'tfg_mysql_lead.csv'
INTO TABLE mydb.lead
FIELDS TERMINATED BY ','
(id_lead,name,@vardate,age,gender,id_region,id_origin)
SET birthday = STR_TO_DATE(@vardate,'%d/%m/%Y %h:%i:%s %p')
```

Se muestran las tablas del modelo de datos en MySQL (izquierda) y las tablas del modelo de datos en Cosmos (derecha)





Se representa el flujograma de importación:



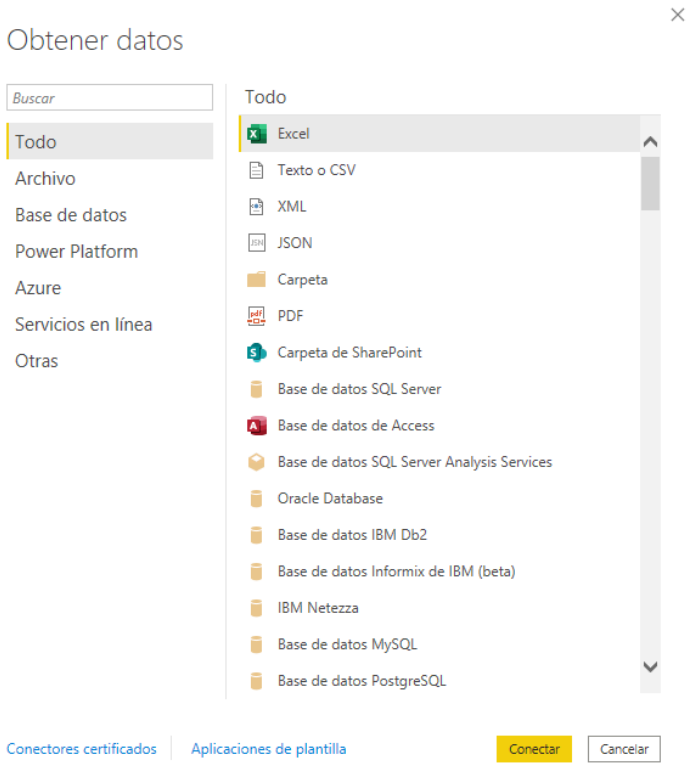
3.3 Interconexión entre MySQL y Power BI

3.3.1 Datos sobre Power BI y su conectividad

Después de indagar sobre las capacidades de los principales software BI, se puede afirmar que Power BI se apoya en tres capacidades para convertirse en uno de los referentes de mercado en soluciones business intelligence: su conectividad, su poder de transformación de datos y su capacidad para visualizar los datos.

Estas tres capacidades cuentan con una ventaja añadida respecto de otras aplicaciones similares, su inmejorable experiencia de usuario que hace su uso algo muy intuitivo y, sobre todo, flexible. Aunque en el mercado existen programas con una mayor implantación en términos históricos como MicroStrategy, SPS o Cognos, Power BI pertenece a esa corriente de fácil usabilidad para personal sin conocimientos de programación como puede ser QlikView, Dundas BI o Tableau.

En este punto del proyecto, nos centramos en las habilidades de interconexión que tiene Power BI para importar datos. Podemos decir que son casi infinitas. Si vamos a la opción “Obtener datos”, encontramos estas secciones:



Vemos que tenemos las siguientes opciones:

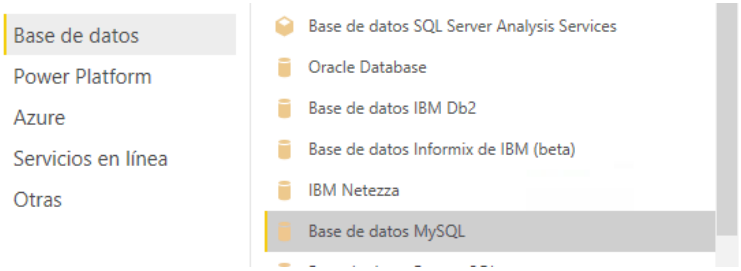
- **Archivo**
  Permite importar datos desde archivos estáticos en los principales formatos como CSV, XML, JSON e incluso desde un archivo PDF.

- **Base de datos**
  Otra fuente de información es la conectividad directa con muchos tipos de base de datos de gran implantación en el mercado como puede ser SQL Server, MySQL, Oracle DB, IBM Db2, PostgreSQL o Sybase.

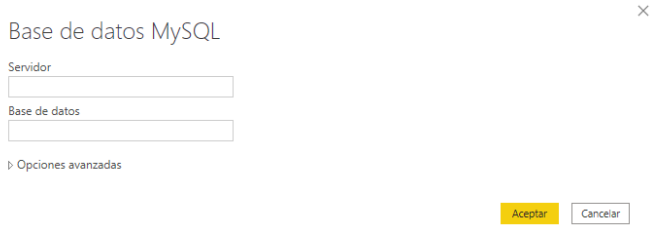
- **Cloud**
  Aquí las opciones son enormes. Desde fuente de datos alojadas en varias utilidades de Azure o Dynamic, hasta opciones muy interesantes como son las conexiones provistas por servicios web o bien software ERP/CRM como Zoho, Webtrends Analytics o directamente un servicio web REST.

3.3.2 Conexión del Data Warehouse con la capa de presentación.

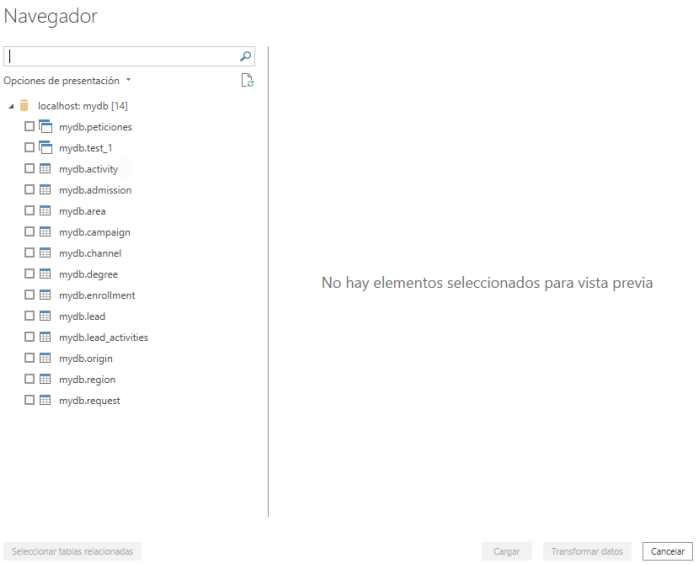
En este proyecto tenemos alojado nuestro Data Warehouse en MySQL. Por lo tanto, procedemos a elegir en la sección “Base de datos” la opción “Base de datos MySQL:



Power BI nos requiere estos datos:



Teniendo en localhost la base de datos creada, procedemos a indicar los datos requeridos. Power BI conecta con MySQL y nos muestra el catálogo de la base de datos. Elegimos todos y ya tenemos acceso a los datos de nuestro Data Warehouse.



Capítulo 4 – Visualización en Power BI

4.1 Estrategia a seguir

Antes de pasar a definir los diferentes *dashboard* o cuadros de mando así como las métricas usadas, hay que aclarar el uso del término *data warehouse* o almacén de datos que se está haciendo en este proyecto. En sentido estricto, un almacén de datos suele seguir un esquema relacional en forma de estrella o incluso, con una elaboración más compleja, un esquema en forma de copo de nieve.

Estos esquemas consisten en tener dos conjuntos de tablas: las tablas de hechos y las tablas de dimensiones. Las tablas de hechos registran medidas relacionadas con un hecho concreto. Por ejemplo, en el proyecto que nos ocupa, podríamos tener una tabla de hechos que registre las admisiones. Las tablas de dimensiones registrarían los atributos que definen esos hechos. En el caso de las admisiones, la fecha no iría en la tabla de hechos, sino que esta contendría una clave foránea vinculada a una tabla de dimensiones que registraría la dimensión temporal. En esta tabla se podría, además, desglosar en diferentes escalas esa fecha: día, mes, año, trimestre, cuatrimestre, etc.

Si bien este es el modelo clásico de almacén de datos, el proyecto actual se ha conducido bajo un modelo relacional que sería el equivalente al que manejaría el software CRM del centro educativo, donde la información se normaliza en entidades principales y maestros con sus respectivos enlaces como se ha definido en el capítulo 2. Sin entrar a detallar las ventajas y desventajas de cada paradigma, vamos a usar este segundo esquema dado que permite evitar un segundo proceso de conversión del modelo relacional de la base de datos del software CRM al modelo en estrella o copo de nieve. Además, vamos a poder realizar una serie de vistas en MySQL que, al actualizar el origen de datos desde Power BI, nos dará un control total de aquellos datos necesarios para nuestros fines. Por supuesto, desde Power BI también podremos, como se verá, añadir nuevas métricas, enlaces y consultas sobre los datos ya importados. Para ello haremos uso de DAX (Data Analysis Expressions), el lenguaje de Power BI para realizar transformaciones y aplicar muchas fórmulas similares, por cierto, a las que se manejan en Excel.

4.2 *Dashboard* sociodemográfico

El objetivo del primer *dashboard* será mostrar al usuario de forma nítida las variables que definen al lead: su ubicación, su género y su edad. Además, se incluirá el origen o estudios previos que permitirá hacerse una idea rápida de cuáles son los diferentes perfiles de alumnos potenciales del centro. A continuación se presenta la definición del *dashboard* atendiendo a los siguientes criterios:

Nombre	Sociodemográfico
Objetivo	Conocer el perfil sociodemográfico del alumno potencial
Tablas implicadas	*lead, region, origin*

Consulta de la vista	`CREATE VIEW sociodemografico AS`
	`SELECT`
	`l.gender lead_genero`
	`, l.age lead_edad`
	`, r.name provincia`
	`, o.name origen, 1 cantidad`
	`FROM`
	`mydb.lead l`
	`, mydb.region r`
	`, mydb.origin o`
	`WHERE`
	`l.id_region = r.id_region`
	`AND l.id_origin = o.id_origin`

Funciones DAX	MAX, MIN, AVERAGE
Elementos gráficos	Circular, Líneas, Barras horizontales, Barras verticales, Medidor

Viendo la consulta SQL, se podría haber realizado una agrupación con la sentencia "group by" de los cuatro campos de la consulta. Sin embargo, se ha decidido incluir una fila por cada registro. Un listado parcial de los datos que devuelve la vista:

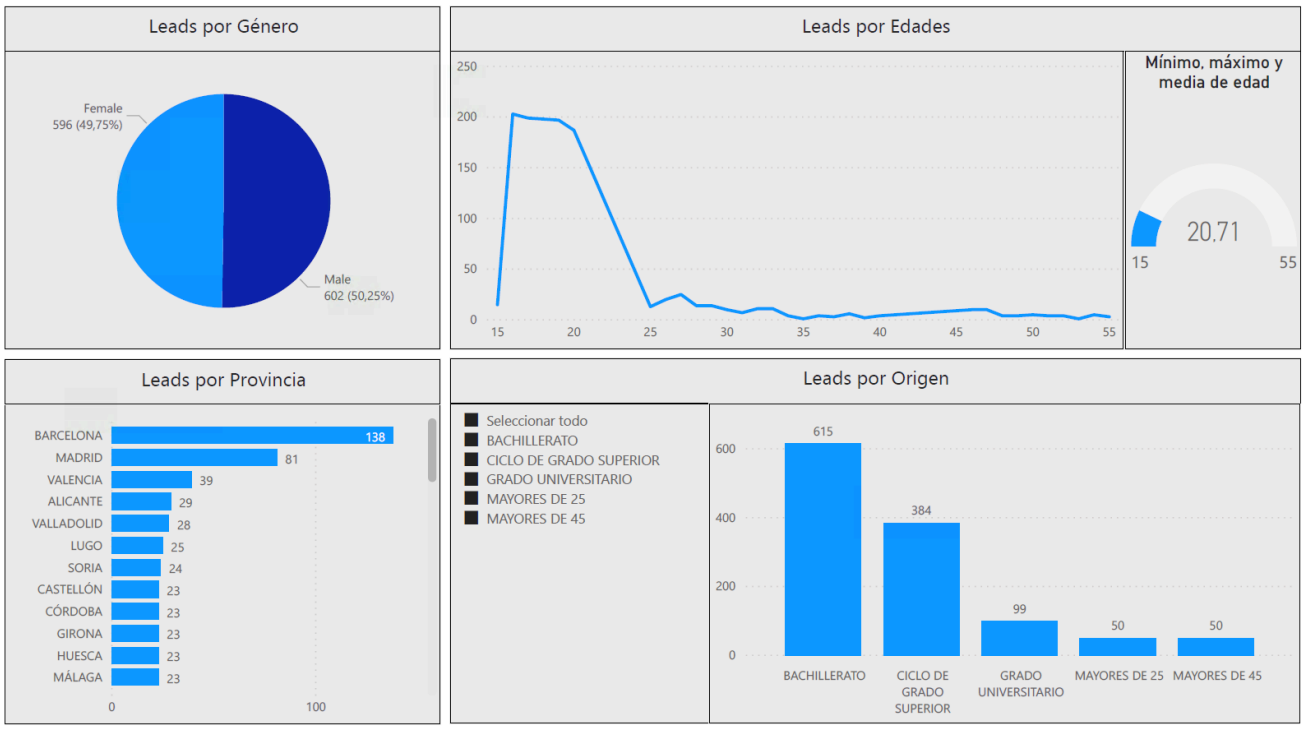
lead_genero	lead_edad	provincia	origen	cantidad
Female	16	BARCELONA	BACHILLERATO	1
Female	16	BARCELONA	BACHILLERATO	1
Male	18	BARCELONA	BACHILLERATO	1
Male	18	BARCELONA	BACHILLERATO	1
Male	17	BARCELONA	BACHILLERATO	1
Male	18	BARCELONA	BACHILLERATO	1
Female	18	BARCELONA	BACHILLERATO	1

Aunque Power BI permite segmentar haciendo clic en los elementos de cualquiera de los gráficos, se ha incluido únicamente un control de segmentación que corresponde con el nivel de estudios previos del lead. Esto permitirá ver la distribución de género, provincia y edades por la selección que, además, se ha habilitado como múltiple, ya que podría ser de interés visualizar, por ejemplo, los mayores de 25 y 45 años y, por otro lado, los bachilleratos y ciclos de grado superior.

Para el género, debido a que sólo se muestran dos posibles valores, se ha optado por un gráfico circular. En el caso de provincias y origen o nivel de estudios previos, se ha creído conveniente gráficos de barras, si bien en el caso de las provincias habría que hacer *scroll* para verlas todas.

La variable edad se ha mostrado de dos formas distintas: Por un lado tenemos un gráfico de líneas con la edad en el eje X y la cantidad de leads en el eje Y. Por otro lado, se visualiza un gráfico de semicírculo medidor con la edad mínima y máxima en ambos extremos. Se rellenará hasta la edad media mostrando los tres valores.

Este es el resultado:



4.3 *Dashboard* académico

El segundo *dashboard* tiene como objetivo mostrar, para las peticiones de los alumnos, los grados agrupados por sus respectivas áreas. Un dato de interés asociado es el género, el cual nos permite ver si existen desviaciones hacia uno u otro por grados solicitados. Veamos el detalle técnico del *dashboard*:

Nombre	Académico
Objetivo	Conocer los grados y áreas de interés atendiendo al género y al origen
Tablas implicadas	*lead, region, origin, request, area, degree*

Consulta de la vista	`CREATE VIEW academico as`
	`SELECT`
	`l.gender lead_genero`
	`, l.age lead_edad`
	`, o.name origen`
	`, a.name area`
	`, d.name titulacion`
	`, 1 cantidad`
	`FROM`
	`mydb.lead l`
	`, mydb.region r`
	`, mydb.origin o`
	`, mydb.request q`
	`, mydb.area a`
	`, mydb.degree d`
	`WHERE`
	`l.id_region = r.id_region`
	`AND l.id_origin = o.id_origin`
	`AND q.id_lead = l.id_lead`
	`AND q.id_degree = d.id_degree`
	`AND d.id_area = a.id_area`

Funciones DAX	GROUP BY
Elementos gráficos	Treemap, Matriz, Segmentadores por botón y slider

Se muestra a continuación el resultado parcial de la consulta de la vista:

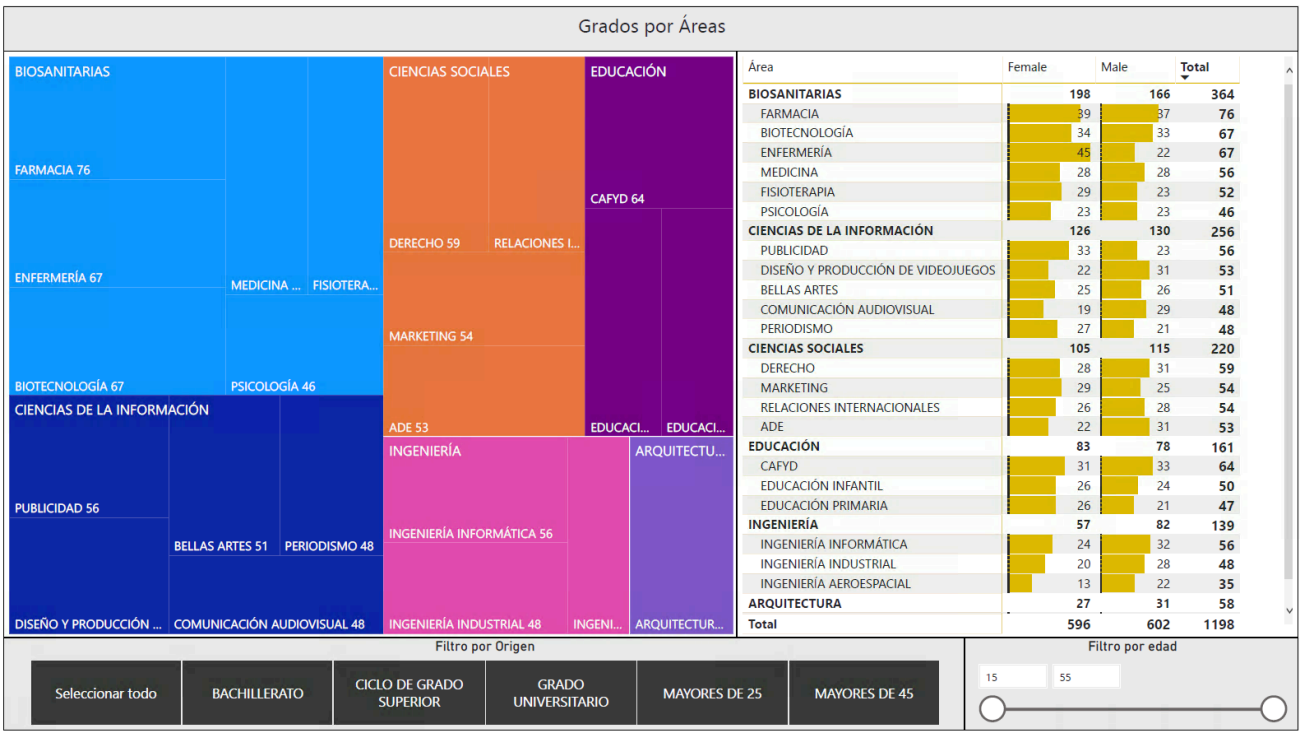
lead_genero	lead_edad	origen	area	titulacion	cantidad
Female	18	BACHILLERATO	CIENCIAS SOCIALES	ADE	1
Female	17	BACHILLERATO	CIENCIAS SOCIALES	ADE	1
Male	17	BACHILLERATO	CIENCIAS SOCIALES	ADE	1
Female	17	BACHILLERATO	CIENCIAS SOCIALES	ADE	1

Para la visualización se podían haber elegido varias opciones como un gráfico donut con dos niveles o un gráfico de barras apiladas. Sin embargo, dado el carácter jerárquico que contiene la relación área y grado, el gráfico Treemap es el más idóneo. Además, con los colores por área, permite comparar de forma sencilla las frecuencias entre las diferentes áreas. Dentro de cada área, el tamaño de cada rectángulo ayuda a hacernos una idea del grado más solicitado y menos solicitado de dicha área. En definitiva, facilita la lectura al mostrar de forma resumida grandes volúmenes de información.

Para complementar el Treemap se hace uso de una matriz con los datos agrupados en las filas y las cantidades desglosadas por género en las columnas con una columna total resaltada en negrita. Un elemento gráfico añadido son las pequeñas barras horizontales dentro de cada celda que permite, sin fijarse en el dato numérico, ver qué grado es el más demandado de cada área.

Finalmente se han incluido dos filtros: Por origen o nivel de estudios previo y por edad. En el caso de la edad, se incluye un slider que ayuda a mejorar la experiencia de usuario.

Este es el resultado:



4.4 *Dashboard* flujo de interacciones

El siguiente *dashboard* consigue visualizar, gracias a un gráfico de flujo, la secuencia de las diferentes interacciones que se dan en el proceso de captación. Existen varios diagramas o gráficos de flujo (también conocidos como gráficos de red) que permiten visualizar este tipo de interrelación entre diferentes elementos. Por ejemplo, el diagrama de cuerdas (*Chord diagram*) es un método interesante que consiste en mostrar en las secciones de un círculo los enlaces entre diferentes elementos con arcos, donde el tamaño de las secciones y de los arcos son las cantidades o frecuencias que hay entre un elemento y otro. Sin embargo, este tipo de gráfico está más orientado a visualizar matrices de datos y no secuencias.

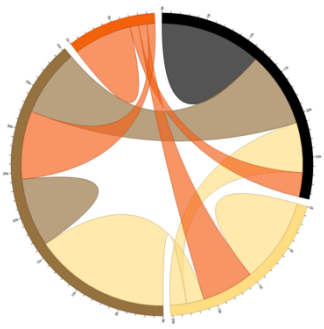


Diagrama de cuerdas paralelas

Similar al gráfico de cuerdas tenemos el gráfico de arcos, donde a diferencia del anterior, en este caso el eje X es lineal y no circular. Además de tener la misma limitación que el gráfico anterior, surge el problema de los cruces entre arcos. Dado nuestro modelo de interacciones, es muy probable que esto suceda y el gráfico no sería nada descriptivo.

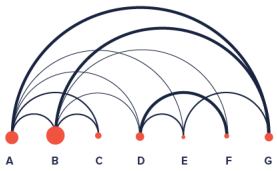


Diagrama de arco

Dentro de los gráficos de flujo o de red, el idóneo para el caso que nos ocupa es el conocido como diagrama de Sankey, ya que no sólo permite visualizar secuencias completas de interacciones con diferentes pasos, sino que el ancho de cada "flecha" muestra la medida de proporcionalidad o, en nuestro caso, la cantidad de interacciones entre un paso y otro del proceso de captación. Además, en las diferentes secuencias de nuestro modelo, pueden existir diferentes pasos intermedios donde se termina la secuencia. Este gráfico también contempla este supuesto.

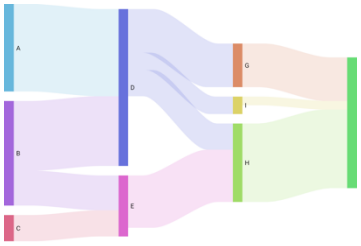


Diagrama de Sankey

Realizar este tipo de gráfico desde Power BI es un reto importante puesto que, depende de cómo se configure el *dataset*, el resultado final puede ser inútil al mostrar sólo los flujos entre diferentes interacciones pero no la secuencia ordenada de dichos flujos.

Para conseguir lo que estamos buscando ha sido necesario añadir una nueva columna a la tabla *lead_activities.* Se trata de la columna *orden*, que irá registrando, para cada interacción de cada lead, el número de orden siendo el valor 1 el de la primera interacción. Ha sido necesario ejecutar un script donde se asigne valor a dicha columna de forma secuencial comenzando por 1, incrementando por cada actividad del mismo lead y reiniciando su valor cuando el lead cambia (el proceso recorre las actividades del lead ordenando por id_lead y el id secuencial de la tabla).

En la siguiente tabla se muestra una consulta de los primeros nueve leads ordenado por id_lead y orden (aunque el resultado es el mismo que si lo ordenamos por el id de la tabla). Se enmarca en recuadro verde el identificador del lead y en recuadro azul el orden para poder ver claramente que el script ejecutado ha asignado correctamente el valor de la nueva columna que nos permitirá ejecutar una query válida para que Power BI visualice correctamente el diagrama Sankey.

id	date	id_previous	orden	id_lead	id_activity	id_campaign
1	2020-02-16	0	1	1	14	1
2	2020-03-07	0	1	2	5	1
3	2020-03-14	2	2	2	1	1
4	2020-03-23	3	3	2	14	1
5	2020-03-11	0	1	3	2	1
6	2020-03-13	5	2	3	7	1
7	2020-03-23	6	3	3	12	1
8	2020-03-13	0	1	4	15	1
9	2020-01-03	0	1	5	4	1
10	2020-01-16	9	2	5	14	1
11	2020-01-24	0	1	6	5	1
12	2020-02-04	11	2	6	1	1
13	2020-02-13	12	3	6	14	1
14	2020-02-02	0	1	7	15	1
15	2020-03-22	0	1	8	4	1
16	2020-04-02	15	2	8	15	1
17	2020-03-20	0	1	9	5	1
18	2020-03-24	17	2	9	1	1

La estrategia en SQL para lograr una consulta válida ha consistido en hacer uso de la función CONCAT que nos permite concatenar cadenas estáticas con columnas para poder indicar el número de paso de cada interacción. También ha sido clave el uso de la columna que almacenaba el identificador de la interacción precedente para poder enlazar una con otra y obtener en una misma fila la interacción completa. Sólo se tienen en cuenta aquellas interacciones donde existe un origen y destino (orden > 1). A continuación se muestra dicha sentencia y el resultado parcial de los datos obtenidos:

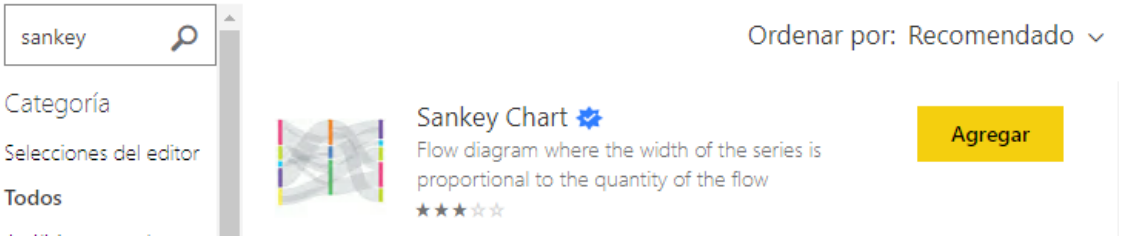
Nombre	Flujo interacciones
Objetivo	Visualizar la secuencia y cantidad de interacciones entre los diferentes pasos que dan los leads en el proceso de captación.
Tablas implicadas	*lead_activities, acitivity, channel*

Consulta de la vista	`CREATE VIEW sankey as`
	`SELECT CONCAT('P',(la.orden-1),"-",a2.name) Origen`
	`, CONCAT("P",(la.orden),"-",a.name) Destino`
	`, count(*) Cantidad`
	`FROM`
	`lead_activities la`
	`, activity a`
	`, channel c`
	`, channel c2`
	`, activity a2`
	`, lead_activities la2`
	`WHERE`
	`la.id_activity = a.id_activity`
	`AND c.id_channel = a.id_channel`
	`AND la2.id_activity = a2.id_activity`
	`AND c2.id_channel = a2.id_channel`
	`AND la.id_previous = la2.id`
	`AND la.orden > 1`
	`GROUP BY 1,2`
	`ORDER BY 1,2`

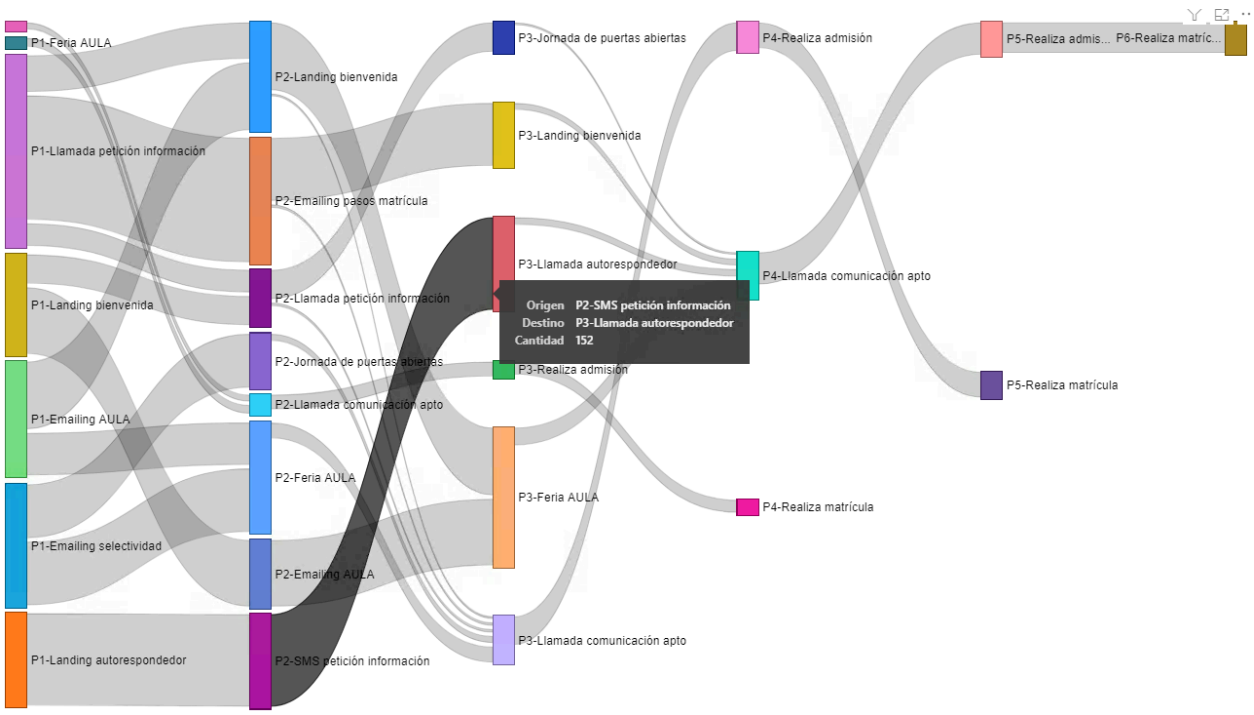
Funciones DAX	(ninguna)
Elementos gráficos	Sankey Diagram

Origen	Destino	Cantidad
P1-Emailing AULA	P2-Feria AULA	69
P1-Emailing AULA	P2-Landing bienvenida	111
P1-Emailing selectividad	P2-Feria AULA	104
P1-Emailing selectividad	P2-Jornada de puertas abiertas	88
P1-Feria AULA	P2-Llamada comunicación apto	13
P1-Jornada de puertas abiertas	P2-Llamada comunicación apto	10
P1-Landing autorespondedor	P2-SMS petición información	152
P1-Landing bienvenida	P2-Emailing AULA	109

Aclarar que para conseguir visualizar este tipo de gráfico es necesario obtenerlo de la galería de objetos visuales de Power BI disponibles ya que no viene en la instalación por defecto.



Después de realizar algunos cambios visuales como el tamaño de la fuente y la eliminación del título, el resultado es el que se muestra a continuación (se ha hecho *mouseover* sobre una interacción para ver el detalle):

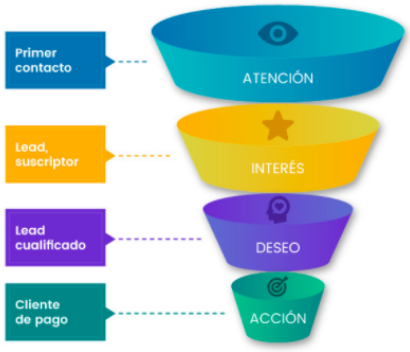


Power BI permite mover los diferentes elementos, pero tal y como le hemos pasado el *dataset*, lo organiza como buscábamos asignando cada sección vertical a uno de los pasos. No se han incluido filtros para obtener una mayor capacidad de visualización en el *dashboard*.

4.5 *Dashboard* Embudo de conversión

Cualquier análisis en un modelo de captación de leads requiere de forma imprescindible lo que se conoce como un embudo de ventas o *funnel chart*. Este tipo de visualización atiende, de forma teórica, a la evolución del lead y su paso por las diferentes fases del proceso de ventas. Actualmente no hay un consenso al respecto y cada empresa lo aplica según la conveniencia a su modelo de negocio y forma de ver el proceso de captación.

A continuación se muestran tres ejemplos que atienden a distintas fases:



Como se puede observar, cada gráfico atiende a criterios distintos que en ocasiones están relacionados con el origen del lead, el sector aplicado o los medios de contacto por los que el lead puede ir formalizando su interés. Sea como fuere, el concepto está claro: El objetivo es establecer una clasificación de interés creciente para ver cómo los leads evolucionan (o no) en el proceso de captación desde que muestra un mínimo interés hasta que finalizan la compra.

En el presente proyecto aplicado al sector educativo, teniendo en cuenta el modelo de datos que se ha desarrollado, vamos a establecer unas fases diferenciadas de evolución del lead atendiendo a dos criterios:

1) Canal por el que se ha realizado el contacto.
2) Número de paso en la secuencia de interacción.

Las fases serán las siguientes:

FASE	DESCRIPCIÓN
1	Pide información
2	Demuestra mayor interés
3	Asiste a evento con contacto previo
4	Realiza admisión
5	Formaliza matrícula

Y las actividades, según canal y número de paso, se encuadrarán en las anteriores fases de la siguiente forma:

ACTIVIDAD	CANAL	PASO	FASE
Landing bienvenida	LANDING PAGE	1	1
Landing autorespondedor	LANDING PAGE	1	1
Landing inscripción AULA	LANDING PAGE	1	1
Llamada petición información	LLAMADA ENTRANTE	1	1
Llamada autorespondedor	LLAMADA SALIENTE	1	1
Emailing selectividad	EMAIL	> 1	2
Emailing AULA	EMAIL	> 1	2
Emailing pasos matrícula	EMAIL	> 1	2
Llamada petición información	LLAMADA ENTRANTE	>1	2
Feria AULA	EVENTO PRESENCIAL	>1	3
Jornada de puertas abiertas	EVENTO PRESENCIAL	>2	3
Realiza admisión	EVENTO PRESENCIAL	>1	4
Realiza matrícula	EVENTO PRESENCIAL	>1	5

Como se puede observar, no todas las actividades se tienen en cuenta a la hora de elaborar el embudo y hay algunas que se contemplan en diferentes fases según el paso en el que se encuentren en la secuencia de interacciones. Estas decisiones son arbitrarias y, en un contexto empresarial, recaerían en el responsable de marketing con el objetivo de afinar el modelo a la par que ajustar las acciones al presupuesto disponible.

Una vez más haremos uso de la potencia que nos brinda SQL para preparar una vista que facilite el trabajo dentro de Power BI. Aun así, como veremos a continuación, hay cierto paso que lo haremos dentro de Power BI con la función SWITCH.

Con SQL la estrategia consiste en hacer uso de la cláusula UNION ALL para sumar todas las filas de las diferentes consultas, una por cada fase del embudo. Eso sí, para poder tener mayor opción de filtrado dentro del dashboard, enlazaremos con otras tablas para obtener el género, el origen y la provincia del lead.

Nombre	Embudo de ventas
Objetivo	Visualizar la conversión de leads en cada fase de las cinco establecidas pudiendo filtrar por género, origen y provincia del lead.
Tablas implicadas	*lead_activities, acitivity, lead, origin, region*

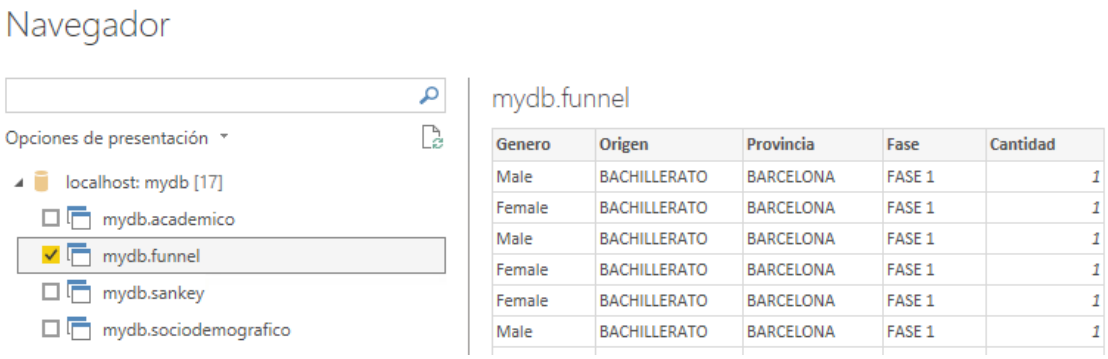
Consulta de la vista	`CREATE VIEW funnel as`
	`SELECT`
	`l.gender Genero`
	`, o.name Origen`
	`, r.name Provincia`
	`,'FASE 1' Fase`
	`, 1 Cantidad`
	`FROM`
	`lead_activities la`
	`, lead l`
	`, origin o`
	`, region r`
	`WHERE`
	`la.id_activity in (1, 2, 3, 10, 12)`
	`AND la.orden = 1`
	`AND la.id_lead = l.id_lead`
	`AND l.id_origin = o.id_origin`
	`AND l.id_region = r.id_region`
	`UNION ALL`
	`SELECT`
	`l.gender Genero`
	`, o.name Origen`
	`, r.name Provincia`
	`,'FASE 2' Fase`
	`, 1 Cantidad`
	`FROM`
	`lead_activities la`
	`, lead l`
	`, origin o`
	`, region r`
	`WHERE`
	`id_activity in (4, 5, 6, 4)`
	`AND orden > 1`
	`AND la.id_lead = l.id_lead`
	`AND l.id_origin = o.id_origin`
	`AND l.id_region = r.id_region`
	`UNION ALL`
	`SELECT`
	`l.gender Genero`
	`, o.name Origen`
	`, r.name Provincia`
	`,'FASE 3' Fase`
	`, 1 Cantidad`
	`FROM`

	lead_activities la , lead l , origin o , region r WHERE id_activity in (14, 15) AND orden > 2 AND la.id_lead = l.id_lead AND l.id_origin = o.id_origin AND l.id_region = r.id_region UNION ALL SELECT l.gender Genero , o.name Origen , r.name Provincia ,'FASE 4' Fase , 1 Cantidad FROM lead_activities la , lead l , origin o , region r WHERE id_activity = 16 AND la.id_lead = l.id_lead AND l.id_origin = o.id_origin AND l.id_region = r.id_region UNION ALL SELECT l.gender Genero , o.name Origen , r.name Provincia ,'FASE 5' Fase , 1 Cantidad FROM lead_activities la , lead l , origin o , region r WHERE id_activity = 17 AND la.id_lead = l.id_lead AND l.id_origin = o.id_origin AND l.id_region = r.id_region
Funciones DAX	SWITCH, TRUE
Elementos gráficos	Funnel Chart, Segmentadores jerárquicos

Se muestra un resultado parcial de la consulta:

Genero	Origen	Provincia	Fase	Cantidad
Male	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Male	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Male	BACHILLERATO	BARCELONA	FASE 1	1
Male	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Male	BACHILLERATO	BARCELONA	FASE 1	1
Female	BACHILLERATO	BARCELONA	FASE 1	1
Male	BACHILLERATO	MADRID	FASE 1	1
Male	BACHILLERATO	MADRID	FASE 1	1
Male	BACHILLERATO	MADRID	FASE 1	1
Female	BACHILLERATO	MADRID	FASE 1	1
Male	BACHILLERATO	MADRID	FASE 1	1

Importamos la vista:



Ahora bien, lo ideal sería mostrar en el gráfico que vamos a elaborar el nombre de la fase para que sea más descriptivo. En este caso vamos a la tabla importada y creamos una nueva columna ("Fase descripción") donde aplicamos la función SWITCH para dar valor a la nueva columna basándonos en la columna "Fase":

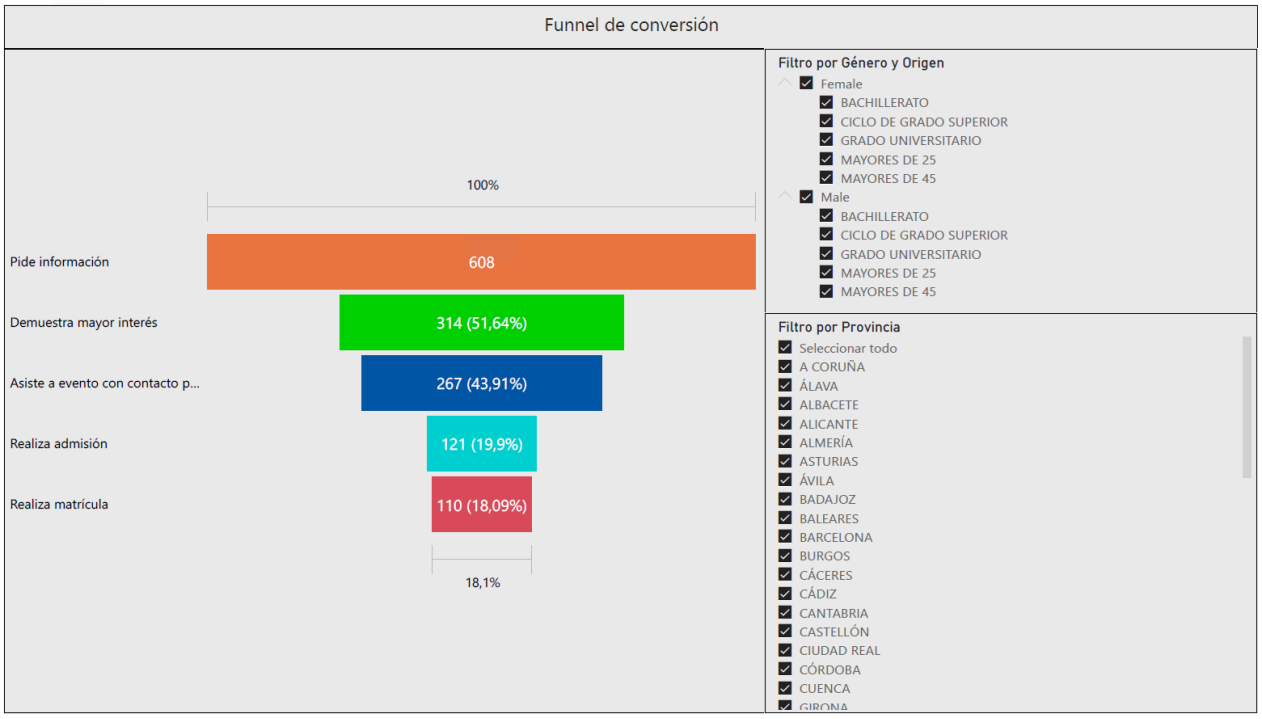
```
Fase descripción = SWITCH(TRUE(),
[Fase]="FASE 1","Pide información",
[Fase]="FASE 2","Demuestra mayor interés",
[Fase]="FASE 3","Asiste a evento con contacto previo",
[Fase]="FASE 4","Realiza admisión",
[Fase]="FASE 5","Realiza matrícula")
```

Y vemos cómo tenemos la nueva columna con los valores descriptivos:

```
1 Fase descripción = SWITCH(TRUE(),
2 [Fase]="FASE 1","Pide información",
3 [Fase]="FASE 2","Demuestra mayor interés",
4 [Fase]="FASE 3","Asiste a evento con contacto previo",
5 [Fase]="FASE 4","Realiza admisión",
6 [Fase]="FASE 5","Realiza matrícula")
```

Genero	Origen	Provincia	Fase	Cantidad	Fase descripción
Male	BACHILLERATO	BARCELONA	FASE 1	1	Pide información
Male	BACHILLERATO	BARCELONA	FASE 1	1	Pide información

Configuramos el *funnel chart* aplicando un color distinto para las diferentes fases e introduciendo los filtros que nos permiten ver cómo afecta cada una de las tres variables en la conversión. Otro estilo aplicado en el *funnel chart* tiene que ver con las etiquetas en cada fase. Se ha optado por incluir un porcentaje de la cantidad de cada fase sobre la primera de todas (Pide información). Vemos, por tanto, que hemos conseguido un ratio del 18,09% de conversión, si bien no estamos midiendo leads, sino aquellas interacciones que hemos decidido tener en cuenta (1420 sobre 3052).



Capítulo 5 – Conclusiones

5.1 Análisis de la planificación y la metodología utilizada

En la sección de planificación se explico el método de división de tareas y gestión de las mismas siguiendo el método kanban. Gracias a su aplicación, ha sido posible estructurar correctamente la dimensión completa del trabajo y hacer un seguimiento visual junto con anotaciones en las diferentes tarjetas.

Respecto de la planificación temporal plasmada en el diagrama de Gantt, se han tenido que realizar diferentes iteraciones en algunos elementos temporales para minimizar los riesgos que se plantearon, especialmente aquellos relacionados con el incumplimiento de timing y objetivos. Para ello, en algunos casos se han realizados saltos hacia tareas futuras con dataset distintos de los finales para asegurarse la capacidad de visualizar los gráficos planeados con unos pequeños dataset completamente aleatorios pero con una estructura igual al modelo de datos. De esta forma, se garantizó que la fase de generación de datos no se retrasaría por tener que repetirla con una estructura distinta.

La fase de implementación final de confección de los dashboard se cumplió en tiempo gracias al uso del lenguaje SQL, mucho más familiar que el uso de las expresiones de análisis de datos (DAX). No considero por ello menor calidad en la implementación ya que, al fin y al cabo, muchas de las funciones DAX se resuelven con mayor sencillez desde SQL con una única sentencia (aunque compleja) que teniendo que lidiar con multitud de transformaciones para alcanzar el mismo resultado.

Finalmente, respecto de las entregas parciales, aquí sí se ha producido una asimetría y cierta desviación contrastando la realidad con la planificación inicial. Las dudas iniciales sobre la temática del proyecto junto con el desconocimiento inicial de la mayoría de frameworks BI, supuso una dilatación en esta fase crítica de la que dependería todo el trabajo. Sin embargo, ese tiempo inicial se recuperó en la fase de implementación de dashboard gracias al sólido modelo de datos y al dominio del lenguaje SQL.

5.2 Evaluación del cumplimiento de los objetivos planteados

1) **Análisis en el modelado de decisiones que pueden influir en los datos extraídos en la fase de implementación final.**

En este sentido, en la fase de modelado se razonó sobre aspectos como contar o no con un maestro para el género e incluir campo de fecha de nacimiento y edad. Además, en el caso del flujo de interacciones, se ha visto cómo incluso planteando una lista enlazada dentro del modelo relacional, para obtener el diagrama de interacciones hizo falta alterar la tabla de interacciones para incluir la columna 'orden' y recalcular de nuevo esa columna. Ha sido un ejemplo práctico de lo que puede suceder con facilidad cuando se crean análisis y se necesitan datos extras a los ya existentes.

2) **Establecer indicadores y técnicas analíticas.**

Finalmente, no se han desarrollado indicadores complejos elaborando modelos estadísticos, aunque sí se han aplicado funciones promedios y establecido convenios como determinar qué acciones correspondían, en función de su orden y repetición, a cada fase del embudo de conversión. Respecto de las técnicas aplicadas, el proyecto se ha apoyado de forma intensiva en el lenguaje SQL para preparar los *dataset* y no tanto en aplicar los recursos que aporta DAX. Y, posteriormente, toda la capa gráfica ha recaído en el uso de Power BI incorporando gráficos extras que están disponibles por defecto en la instalación estándar.

3) **Exposición teórica e implementación gráfica con recursos no usados habitualmente.**

Este objetivo se ha cumplido de forma concreta con el diagrama *Sankey*. Se han expuestos varios diagramas de flujo y se ha razonado la elección de este tipo de diagrama como el más idóneo para representar algo tan complejo como el flujo de interacciones.

5.3 Líneas de trabajo futuro

Es sorprendente cuánta información se ha podido obtener para analizar el perfil del alumno potencial, su interés y su evolución en el proceso de venta con un modelo de datos de apenas doce tablas. Sin embargo, todavía se podría haber obtenido más conocimiento. Un apartado muy relevante es el coste que supone conseguir un lead. Cuando se diseñó el modelo, se incluyó un coste estimado por canal. Las acciones tenían un canal y, por tanto, se podría evaluar el coste de cada camino que sigue el lead, del paso por cada fase del embudo de conversión, etc.

Otro dato que contiene el modelo y que tiene un potencial importante es la fecha de la interacción. Al tener la trazabilidad completa de todas las interacciones del lead con el centro educativo, sería posible calcular cuánto tarda en responder un lead a una acción así como los tiempos entre acciones.

Si se unen estos dos aspectos: coste y tiempo junto con la agrupación en campañas de diferentes secuencias de acciones, el departamento de marketing podría ahondar en los siguientes estudios:

1) Calcular qué acciones son más reactivas para la evolución de un lead.
2) Medir qué acciones o secuencia de acciones derivan en una mayor conversión a un menor coste y qué acciones dejan al lead huérfano en una fase temprana del embudo de conversión.
3) Conocer la rentabilidad de diferentes campañas.
4) Planificar en tiempo las campañas para que coincidan con el periodo de admisión y matriculación en aquel momento en el que suelen provocar mayor conversión en el lead.
5) Descartar acciones de alto coste con un baja conversión.

El presente proyecto tampoco ha utilizado, dado el límite de extensión del proyecto, con la variable geográfica del lead, Y no tanto a nivel provincia, sino a nivel código postal o, más concretamente, sección censal. El cruce de datos del INE con este dato geográfico del lead permite diferenciar clústeres en función del nivel económico que ayudaría a enfocar las acciones con mayor porcentaje de acierto. Especialmente cuando se trata de centros privados de alto poder adquisitivo. Obteniendo la dirección del lead se pueden deducir estos datos e incorporar un nuevo dashboard geográfico con mayor detalle. Si contrastamos esta capa con la capa de los actuales alumnos, se podrían categorizar diferentes zonas con mayor garantía de conversión del lead.

Otra línea futura de actuación tiene que ver con ampliar el modelo de datos para tener en cuenta más información del lead. Por ejemplo, se sabe que un perfil de alumno potencial con un elevado ratio de conversión es aquel conocido en el sector como "alumni" o antiguo alumno. No sólo por el potencial de que curse máster u otro grado, sino por la influencia positiva que puede tener en familiares y amigos. Incorporar esta variable a los lead para saber que es alumni o bien que viene recomendado por un alumni puede darnos no sólo la capacidad de realizar campañas especiales, sino trazar una red de contactos que permitan dirigir acciones muy específicas para este colectivo.

Por otro lado, en el modelo del proyecto no se ha ahondado en la dimensión que las redes sociales pueden aportar a todo el conocimiento derivado del lead. El sector educativo tiene como particularidad que su objetivo perfil es gente joven e incluso adolescentes. Si se recogiera, bajo consentimiento del lead, información de su red social así como permisos para ponerse en contacto por dicha red, esto no sólo aportaría una fuente importante de información, sino que permitiría incorporar nuevas acciones de marketing y "jugar" con ellas dentro de las secuencias que formarían con canales más clásicos como la llamada telefónica o el envío de mensajes SMS.

Existe un factor que abre toda una nueva capa de análisis del lead (y muy relacionada con los eventos presenciales) que sería el centro educativo previo del alumno. Si incorporamos ese dato, el departamento de marketing contaría con una nueva variable que podría indicar mucho acerca de la conversión.

Finalmente, una extensión del modelo que se hace necesaria dada la normativa de europea de protección de datos RGPD (Reglamento General de Protección de Datos), es la incorporación de los padres como leads. No sólo por el hecho de que este sector maneja información de menores de edad, sino por la influencia que pueden tener los progenitores en la decisión de cursar en el centro determinado plan de estudios. Sin embargo, esto es un aspecto más estratégico que apenas aportaría información analítica. Pero, dado el sector elegido, es necesario dejar constancia de un factor tan crítico a nivel legal.

Capítulo 6 – Glosario

- **Lead**: Traducido al español como prospecto, se trata de aquel cliente potencial que ha mostrado interés en algún producto o servicio de una empresa de forma voluntaria y tangible a través de algún canal de comunicación como puede ser un formulario web, un teléfono de contacto o un email.

- **Landing page**: Traducido al español como página de aterrizaje, es una página web de diseño sencillo con un formulario para solicitar información. Suele ser el objetivo de campañas promocionales para dirigir a los leads a un punto donde puedan ponerse en contacto con la empresa de forma rápida.

- **Emailing**: Técnica de marketing que consiste en el envío de correos electrónicos a aquellos contactos que nos han cedido su consentimiento con el objetivo de captar la atención para informar, vender o cualquier otra acción que se considere oportuna. Gracias al seguimiento sobre el email, suele aplicarse sobre el resultado del envío diferentes métricas que indican el éxito de cada acción concreta basándose en las aperturas, clics y otros estados del email enviado.

- **Conversión**: Acción realizada por un cliente potencial que supone el avance en las fases que conducen a su transformación como cliente final.

- **CSV**: Comma-separated values. Es un formato de ficheros que almacena los datos estableciendo el convenio de separar las filas con saltos de línea, las columnas con comas y, en ocasiones, entrecomillar el contenido de cada dato. Existe un estándar validado por IETF (Internet Engineering Task Force) bajo el documento RFC 4180.

- **Dataset**: Es un conjunto de datos que suele estar estructurado con algún formato de fichero o esquema que permite diferenciar filas de columnas para identificar unívocamente cada dato.

- **Data Warehouse**: Traducido como almacén de datos, es la base de datos de una organización donde se integran todos los datos provenientes de diferentes fuentes y que sirven para aplicar técnicas de análisis que permitan obtener información y conocimiento. Es uno de los componentes más importantes de la estrategia de inteligencia de negocio.

- **ETL**: Extract, Transform and Load. Se define con este acrónimo al proceso que permite manipular uno o varios orígenes de datos para su formateo, limpieza, transformación y consolidación adecuándose al diseño del almacén de datos.

- **SQL**: *Structure Query Language*. Es el lenguaje de definición, manipulación y control de datos para manejar sistemas de gestión de bases de datos en esquema relacional. Existe un estándar ANSI aunque diferentes fabricantes de motores de bases de datos han realizado sus propias especificaciones y extensiones como Oracle, MySQL o Postgre entre otros.

- **Query**: Traducido al español: consulta. Se aplica para cualquier instrucción realizada en algún lenguaje formal, como puede ser SQL, para la obtención de un subconjunto de datos que cumplen una serie de condiciones.

- **DAX**: *Data Analysis Expressions*. Es un lenguaje de consultas que utilizan diferentes herramientas como Power BI y que incorpora multitud de funciones para realizar los cálculos, filtros y transformaciones necesarias que permiten obtener los datos adecuados para su posterior análisis.

- **Dashboard**: Panel de control. Concretamente en Power BI se trata de un espacio o pantalla única donde se pueden ubicar diferentes elementos gráficos que resumen información y, normalmente, permite cierta manipulación especialmente para filtrar o segmentar.

- **Kanban**: Método de planificación y seguimiento para gestionar las tareas que componen un trabajo consistente en el uso de tableros que representan diferentes estados de una tarea. Se originó en la década de los 50 en la empresa Toyota y recientemente, gracias a las metodologías ágiles, se ha popularizado en otros sectores, especialmente en el desarrollo de software.

- **Funnel chart**: Gráfico de embudo que permite visualizar la conversión de los leads por cada fase del proceso de venta.

- **Diagrama de Sankey**: Tipo concreto de diagrama de flujo que permite visualizar transferencias de unidades (energía, dinero, etc.) entre unos elementos u otros ya sea de ganancia o pérdida entre ellos pudiendo contemplar la secuencia temporal de dichos flujos.

Capítulo 7 – Bibliografía

- Marketing

  - ¿Qué son los leads y por qué son tan importantes en el Marketing Digital? https://rockcontent.com/es/blog/leads-1/
  - ¿Qué es el emailing y cuáles son las buenas prácticas a seguir? https://es.sendinblue.com/blog/que-es-el-emailing/
  - ¿Qué es una landing page? https://www.40defiebre.com/que-es/landing-page
  - Embudo de ventas https://es.wikipedia.org/wiki/Embudo_de_ventas

- ETL

  - Common Format and MIME Type for Comma-Separated Values Files https://tools.ietf.org/html/rfc4180
  - Data Warehousing, Data Warehouse y Datamart http://josepcurto.com/2007/10/07/data-warehousing-data-warehouse-y-datamart/
  - Códigos de provincia – España https://es.geopostcodes.com/Spain_Provinces

- Creación de dashboard

  - Conexión Power BI con MySQL https://docs.microsoft.com/es-es/power-query/connectors/mysqldatabase
  - Power BI (Referencia general) https://powerbi.microsoft.com/es-es/
  - Sankey Definitions http://www.sankey-diagrams.com/sankey-definitions/
  - Creación y uso de gráficos de embudo en Power BI https://docs.microsoft.com/es-es/power-bi/visuals/power-bi-visualization-funnel-charts

Capítulo 8 – Anexos

En la siguiente dirección web se puede acceder, con los permisos que se han concedido al consultor, a los informes en el entorno Power BI. También se podrían conceder permisos a usuarios de la organización UOC.

https://app.powerbi.com/groups/me/reports/d84ae3a4-3fc7-4718-b22b-37e074165a7c?ctid=aec762e4-3d54-495e-a8fe-4287dce6fe69