Universitat Oberta de Catalunya

uoc.edu

Diseño e implementación de un sistema analítico para datos de Open Payments

Raúl Caro Moreno
Máster Inteligencia de Negocio y Big Data
Itinerario Big data y sistemas NoSQL

Francesc Julbe López (Consultor)
Josep Curto Díaz (Profesor responsable de la asignatura)

25/01/2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Diseño e implementación de un sistema analítico para datos de Open Payments*
Nombre del autor:	*Raúl Caro Moreno*
Nombre del consultor/a:	*Francesc Julbe López*
Nombre del PRA:	*Josep Curto Díaz*
Fecha de entrega (mm/aaaa):	01/2019
Titulación:	*Máster Inteligencia de Negocio y Big Data*
Área del Trabajo Final:	*TFM Big Data*
Idioma del trabajo:	*Español*
Palabras clave:	*open data, big data, business intelligence*

Resumen del Trabajo:

La relación entre la industria de tecnologías sanitarias y los sistemas sanitarios nacionales es cada vez más colaborativa, por lo que es fundamental cumplir unos criterios éticos y de transparencia. En este sentido, Open Payments surge como una iniciativa open data que publica libremente los datos relativos a los pagos que la industria de tecnologías sanitarias realiza a médicos y hospitales.

En este contexto, el presente trabajo ha abordado el diseño e implementación de un sistema de explotación de los datos de Open Payments que permita extraer la información necesaria para responder varias preguntas analíticas planteadas. Además, este sistema permite la ampliación a cualquier análisis que se plantee en el futuro.

El diseño se ha basado en componentes open source con una amplia comunidad de usuarios, lo que elimina el coste de licencias y facilita el desarrollo. La implementación se ha hecho en un ordenador personal pero la arquitectura implementada se puede migrar fácilmente a una infraestructura productiva que permita trabajar a varios equipos simultáneamente.

El desarrollo del trabajo ha seguido fundamentalmente una metodología en cascada, aunque alguna fase ha adoptado una metodología iterativa más cercana a las metodologías ágiles.

El resultado del trabajo ha sido el sistema de explotación y las respuestas analíticas esperadas. Se ha concluido que puede ser complicado considerar un escenario como big data. Además, ha quedado patente que los diferentes interesados (expertos del negocio, analistas, especialistas en sistemas TI) deben trabajar coordinadamente para lograr el éxito en los proyectos.

Abstract:

The relationship between the health technology industry and the national healthcare systems is becoming more and more collaborative, so it is essential to accomplish with the ethical and transparency criteria. In this sense, Open Payments emerges as an open data initiative that publishes data related to health technology industry's payments to physicians and hospitals.

In this context, the present work has performed the design and implementation of an operation system of Open Payments data that requires the necessary information to answer several analytical questions. In addition, this system allows to expand any analysis that may arise in the future.

The design has been based on open source components with a wide community, which eliminates the cost of licenses and facilitates development. The implementation has been done on a personal computer but the implemented architecture can be easily migrated to a productive infrastructure that allows several current systems to work.

The development of the work has fundamentally followed a waterfall methodology, although some phase has used an iterative methodology that is closer to agile methodologies.

The result of the work has been the expected operation system and the analytical responses. It has been concluded that it can be complicated to decide if the use case must be considered a big data scenario. In addition, we have shown that the different stakeholders (business experts, analysts, specialists in IT systems) must work in coordination to achieve success in the projects.

Índice

1. Introducción 1
1.1. Contexto y justificación del Trabajo 1
1.2. Objetivos del Trabajo 1
1.3. Enfoque y método seguido 2
1.4. Planificación del Trabajo 2
1.5. Breve sumario de productos obtenidos 3
1.6. Breve descripción del resto de capítulos de la memoria 4
2. Análisis de los datos disponibles 4
2.1. Análisis preliminar de los datos 4
2.2. Preguntas analíticas a resolver 5
3. Diseño del sistema 5
3.1. Análisis de las características de los datos ingestados 5
3.2. Arquitectura del sistema global 5
3.3. Diseño del sistema de ingesta 6
3.4. Diseño de modelo de datos 7
3.5. Diseño del sistema de almacenamiento 9
3.6. Diseño de los procesos de ETL 10
3.7. Diseño del sistema de análisis 10
3.8. Diseño de la visualización de los datos analizados 10
4. Implementación del sistema 11
4.1. Implementación del sistema de almacenamiento 11
4.2. Implementación de los procesos de ETL 12
4.3. Implementación del sistema de análisis 14
4.4. Implementación de la visualización de los datos analizados 16
5. Respuesta a las preguntas analíticas 16
5.1. ¿Qué tipos de pagos son los más repetidos? 16

5.2. ¿Qué países y estados son los más influenciados? 18
5.3. ¿Qué receptores e investigadores son los más influenciados? 21
5.4. ¿Cuáles son las terceras partes más beneficiadas de los pagos? ¿Tienen las aportaciones a caridad un peso importante en el programa? 22
5.5. ¿Cuáles son los principales destinos de viajes pagados por el programa? 25
5.6. ¿Qué relación tienen los intereses de los médicos en los pagadores y los pagos recibidos? 26
6. Conclusiones 28
7. Bibliografía 30
Anexos 32
8. Diagrama de Gantt de las tareas 32
9. Open Payments Methodology Overview & Data Dictionary 33
10. Catalogación de los datos de Open Payments 34
11. Script de implementación del modelo de datos 43
12. Formatos de las conexiones con la base de datos y de los CSV en Talend Open Studio 43
13. Trabajos de los procesos de ETL 43
14. Script de fragmentación de ficheros CSV 44
15. Tiempos de carga inicial de datos 45
16. Script de carga de las tablas simplificadas de h_payment 47
17. Script de cálculo de indicadores 47
18. Tableros de Qlik Sense 48
18.1. Análisis por tipo de pago 48
18.2. Análisis por localización 49
18.3. Ranking por entidades 50
18.4. Análisis de terceras partes 51
18.5. Análisis de los viajes 52
18.6. Análisis de los intereses de los médicos 53

Lista de figuras

Figura 1 - Arquitectura preliminar del sistema analítico 6

Figura 2 - Modelo de datos 8

Figura 3 - Ejemplo de estructura de subtrabajos de la ETL 13

Figura 4 - Número de pagos, importe total e importe medio por pago en función del tipo de receptor y del año 17

Figura 5 - Distribución de la forma y la naturaleza de los pagos por año 18

Figura 6 - Distribución de la forma y la naturaleza de los pagos en 2017 18

Figura 7 - Distribución del importe de los pagos entre 2013 y 2017 19

Figura 8 - Distribución del importe de los pagos fuera de EEUU por país (de izquierda a derecha y de arriba abajo, total, hospitales, entidades no cubiertas y médicos) 20

Figura 9 - Distribución del importe de los pagos en EEUU por estados (de izquierda a derecha, total y hospitales) 20

Figura 10 - Distribución del importe de los pagos en EEUU por estados (de izquierda a derecha, entidades no cubiertas y médicos) 21

Figura 11 - Importe y proporción de los pagos a terceras partes destinado a la caridad (a la izquierda, hospitales; a la derecha, médicos) 23

Figura 12 - Proporción de los pagos a terceras partes destinado a la caridad (de izquierda a derecha, 2013, 2014, 2015, 2016 y 2017) 24

Figura 13 - Distribución de las aportaciones a caridad en EEUU por estados (de izquierda a derecha y de arriba abajo, 2013, 2014, 2015, 2016 y 2017) 24

Figura 14 - Distribución de los pagos asociados a viajes por país (izquierda) y por estado de EEUU (derecha) 25

Figura 15 - Distribución de los pagos asociados a viajes fuera de Estados Unidos por país 26

Figura 16 - Distribución de los pagos asociados a viajes por ciudad 26

Figura 17 - Indicadores numéricos relativos al análisis de los intereses de los médicos 27
Figura 18 - Gráfico de dispersión de importes recibidos vs cantidades invertidas 28
Figura 19 - Planificación de las tareas del TFM 32
Figura 20 - Análisis por tipo de pago 48
Figura 21 - Análisis por localización 49
Figura 22 - Ranking por entidades 50
Figura 23 - Análisis de terceras partes 51
Figura 24 - Análisis de los viajes 52
Figura 25 - Análisis de los intereses de los médicos 53

Lista de tablas

Tabla 1 - Top 3 de los países que reciben más pagos 19

Tabla 2 - Top 3 de los países que reciben mayores importes 19

Tabla 3 - Top 5 de los hospitales que reciben mayor importe en pagos 22

Tabla 4 - Top 5 de las entidades no cubiertas que reciben mayor importe en pagos 22

Tabla 5 - Top 5 de los médicos que reciben mayor importe en pagos 22

Tabla 6 - Top 5 de los investigadores implicados en mayor importe 22

Tabla 7 - Top 5 de las terceras partes que han recibido mayor importe no asociado a caridad 24

Tabla 8 - Top 5 de las terceras partes que han recibido mayor importe asociado a caridad..... 24

Tabla 9 - Top 10 de las ciudades con mayor importe dedicado a viajes 26

Tabla 10 - Catalogación de los datos de Open Payments 42

Tabla 11 - Tiempos de carga inicial de los datos de Open Payments 46

1. Introducción

1.1. Contexto y justificación del Trabajo

El desarrollo de la industria de tecnologías sanitarias (farmacéuticas y de instrumental médico) siempre ha estado ligada a los sistemas sanitarios nacionales. Sin embargo, en las últimas décadas, se ha pasado de una relación proveedor-cliente a una relación más colaborativa.

En un informe de la consultora PwC de 2013 [1], se ilustra cómo un futuro saludable del sector pasa por la colaboración entre la industria de tecnologías sanitarias y los sistemas sanitarios nacionales. Sin embargo, esta colaboración debe cumplir unos criterios éticos y de transparencia, sobre todo en el ámbito público. Para garantizar estos criterios, se establecen mecanismos como el programa Open Payments o el código de buenas prácticas de la industria farmacéutica española [2].

Open Payments es un programa nacional de transparencia de los EEUU que recopila y publica información relativa a relaciones financieras entre la industria de la salud (por ejemplo, compañías farmacéuticas y de equipamiento médico) y proveedores de servicios de salud (por ejemplo, médicos y hospitales universitarios) con el objetivo de dotar de mayor transparencia al sistema de salud estadounidense [3].

En este contexto, el presente trabajo se plantea explotar los datos de Open Payments para poder analizar las relaciones entre la industria y los sistemas de salud.

1.2. Objetivos del Trabajo

La finalidad de este trabajo es el diseño e implementación de un sistema de inteligencia de negocio / big data que permita la explotación de los datos de Open Payments. Esta finalidad a alto nivel se puede desglosar en los siguientes objetivos concretos:

1. Comprender que tipología de datos ofrece el portal.
2. Identificar cual es la arquitectura más adecuada para representar las relaciones subyacentes.
3. Proponer un objetivo de análisis teniendo en cuenta los datos disponibles en el portal.
4. Determinar cuál es la mejor forma de conseguir datos del portal.
5. Determinar la viabilidad de uso de los datos.
6. Identificar que los diferentes conjuntos de datos son integrables en el marco del objetivo del proyecto.
7. Identificar otras fuentes de datos interesantes para el proyecto.

8. Identificar la forma óptima de extracción de datos de las diferentes fuentes.
9. Adquirir los datos de las diferentes fuentes de datos
10. Diseñar un sistema de inteligencia de negocio / big data que permita almacenar la información adquirida.
11. Implementar este sistema de inteligencia de negocio / big data de forma que permita la ingestión, el almacenamiento, el procesamiento y el análisis de los datos y la visualización de la información obtenida.
12. Elegir los componentes que se usaran para el sistema de inteligencia de negocio / big data de forma que se cubran todas las necesidades del proyecto.
13. Revisar el estado de arte de los componentes seleccionados para el proyecto.

1.3. Enfoque y método seguido

La estrategia para abordar el presente trabajo consiste en trabajar los siguientes bloques:

- Plantear las preguntas analíticas que queremos contestar. Estas preguntas condicionarán el resto de bloques
- Analizar los datos disponibles para diseñar una arquitectura y un modelado de datos de datos adecuado.
- Establecer los mecanismos de ingesta de datos adecuados.
- Diseñar e implementar el sistema analítico adecuado para resolver las preguntas planteadas.
- Definir la manera más adecuada de presentar los resultados del sistema analítico.
- Responder las preguntas planteadas a partir de la información obtenida.

Esta estrategia, ejecutada de forma secuencial, permite disponer de todos los elementos necesarios al inicio de cada bloque de trabajo.

1.4. Planificación del Trabajo

A continuación, se desglosan las tareas a realizar:

- Análisis de los datos disponibles.
    - Analizar los datos de Open Payments (tipos, contenido, etc.).
    - Plantear las preguntas analíticas a resolver.
    - Diseñar arquitectura/modelo de datos.
    - Diseñar estrategia de obtención de los datos de Open Payments.

    - Identificar otras fuentes de datos útiles.
    - Diseñar estrategia de obtención de las otras fuentes de datos.
- Diseño del sistema de inteligencia de negocio / big data.
    - Analizar las características de los datos ingestados.
    - Diseñar el sistema de ingesta.
    - Diseñar el sistema de almacenamiento.
    - Diseñar los procesos de ETL.
    - Diseñar los procesos de análisis.
    - Diseñar el sistema de visualización de los datos analizados.
- Implementación del sistema inteligencia de negocio / big data.
- Contestar las preguntas analíticas a partir de la información obtenida.
- Finalización y entrega de la memoria y la presentación.

En el diagrama de Gantt representado en el Anexo 8, se muestra la planificación temporal de las tareas. Los hitos que se irán alcanzando estarán asociados a los siguientes entregables:

- H1: Análisis de los datos disponibles → 28-11-2018
- H2: Diseño del sistema de inteligencia de negocio / big data → → 27-12-2018
- H3: Implementación del sistema de inteligencia de negocio / big data → → 11-01-2019
- H4: Respuesta a las preguntas analíticas → 18-01-2019
- H5: Memoria y presentación finales del trabajo → 25-01-2019

1.5. Breve sumario de productos obtenidos

El producto resultante de la realización del presente trabajo será un sistema analítico integral que permita responder preguntas analíticas relativas a Open Payments. Este producto se puede descomponer en varios subproductos:

- Un sistema de almacenamiento de los datos ingestados que permita el procesamiento y análisis de dichos datos para obtener la información que permita responder las preguntas analíticas planteadas.
- Un sistema de visualización de la información analizada que permita interpretar la información para poder responder las preguntas analíticas planteadas.

1.6. Breve descripción del resto de capítulos de la memoria

A continuación, se describen los diferentes capítulos de la memoria.

En el capítulo 2, se elabora un análisis preliminar de los datos disponibles y se plantean las preguntas analíticas que se espera responder usando dichos datos.

En el capítulo 3, se realiza el diseño completo del sistema propuesto. Este diseño consistirá en la definición de los procesos de ingesta, la creación del modelo de datos y la elección de las herramientas adecuadas para los diferentes subsistemas.

En el capítulo 4, se explica el proceso de implementación de los diferentes subsistemas que componen la solución propuesta. Abarca la instalación y configuración de las herramientas, la generación de los procesos específicos que requiere el proyecto y la justificación de las decisiones tomadas.

En el capítulo 5, se detalla cómo se han procesado los datos, cómo se presenta la información obtenida y se interpreta dicha información respondiendo a las diferentes preguntas analíticas planteadas.

En el capítulo 6, se exponen las conclusiones relacionadas con la elaboración de este trabajo y se comentan los aspectos que no se han podido abordar y que sería interesante acometer en próximas fases.

2. Análisis de los datos disponibles

2.1. Análisis preliminar de los datos

El primer paso ha consistido en descargar los datos disponibles en la página web de Open Payments [4] y analizar su contenido y su estructura. Open Payments pone a disposición de los usuarios una guía de metodología / diccionario de datos [5] donde se explica la cómo se recopilan los datos y, lo que es más importante, su estructura y el significado de los diferentes campos para los diferentes años de publicación. En el Anexo 9, se adjunta un resumen de elaboración propia a partir de dicha guía.

A partir del análisis de la guía de metodología / diccionario de datos [5], se observa que los datos que proporciona Open Payments son relativos a:

- Los pagos efectuados (a beneficiarios del programa o a terceros).
- La publicación de los datos (fecha, disputas, etc.).
- El pagador y receptor del pago (médicos, investigadores, hospitales).
- Los productos o viajes que intervienen en las transacciones.
- Los intereses del receptor de los pagos en la entidad que los realiza.
- Las investigaciones relacionadas con los pagos.

2.2. Preguntas analíticas a resolver

A partir de los datos disponibles, las preguntas analíticas que deberá resolver el sistema analítico a desarrollar son las siguientes.

- ¿Qué tipos de pagos son los más repetidos?
- ¿Qué países y estados son los más influenciados?
- ¿Qué receptores e investigadores son los más influenciados?
- ¿Cuáles son las terceras partes más beneficiadas de los pagos? ¿Tienen las aportaciones a caridad un peso importante en el programa?
- ¿Cuáles son los principales destinos de viajes pagados por el programa?
- ¿Qué relación tienen los intereses de los médicos en los pagadores y los pagos recibidos?

Los datos de Open Payments son suficientes para responder estas preguntas, por lo que no es necesario buscar fuentes de datos adicionales.

3. Diseño del sistema

3.1. Análisis de las características de los datos ingestados

Antes de trabajar en la arquitectura del sistema, se analizan las características de los datos que van a alimentar el sistema analítico. Desde el punto de vista de las tres V:

- Velocidad: Los ficheros que se ingestan en la plataforma se publican una vez al año, por lo que la velocidad no justifica la consideración de big data.
- Variedad: Cada año se publican únicamente tres ficheros con una estructura fija, por lo que la variedad no justifica la consideración de big data.
- Volumen: Los ficheros publicados anualmente tienen unos 12 millones de registros y ocupan unos 6.5GB, por lo que el volumen no justifica la consideración de big data.

En conclusión, no se deben tener en cuenta criterios de big data para diseñar el sistema de analítica de datos de Open Payments. De todos modos, el volumen acumulado de datos entre 2013 y 2017 es considerable (unos 30GB en total).

3.2. Arquitectura del sistema global

En la Figura 1, se presenta el esquema de la arquitectura a alto nivel del sistema global.

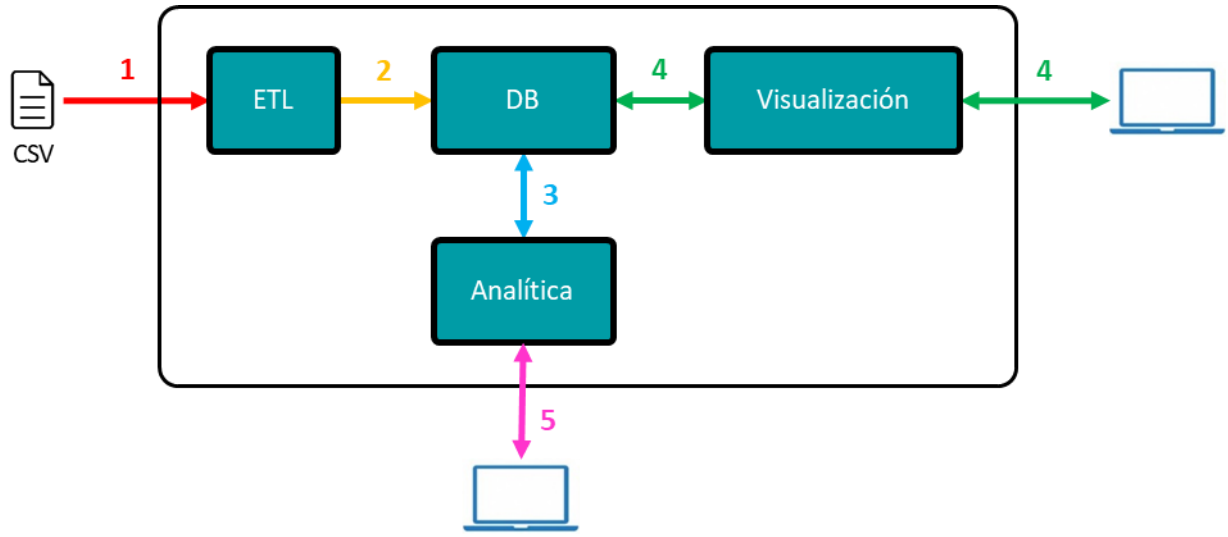


Figura 1 - Arquitectura preliminar del sistema analítico

A continuación, se describe el camino de los datos:

1. El administrador del sistema comprobará la publicación de los datos en junio de cada año y verificará que se mantiene la estructura esperada. Si todo es correcto, lanzará los procesos de ETL y verificará que el proceso ha sido correcto. En caso de que haya algún problema se dará marcha atrás, recuperando los datos anteriores a la actualización y se corregirá el problema.
2. El proceso de ETL cargará los datos en la base de datos, quedando a disposición para los procesos de analítica.
3. El sistema analítico recuperará los datos necesarios de la base de datos, los procesará y guardará los resultados en la base de datos. De este modo, estarán disponibles para la visualización.
4. El sistema de visualización recuperará los datos previamente calculados y los mostrará de manera amigable a los usuarios finales.
5. El sistema analítico permitirá a los usuarios analíticos desarrollar los procesos de análisis y calcular los indicadores que se considere oportunos.

3.3. Diseño del sistema de ingesta

El hecho de que la publicación de los datos de Open Payments sea anual permite que el inicio de la ingesta sea gestionado por un humano. Además, no hay garantía de que, en un año determinado, haya cambios de estructura que obliguen a hacer ajustes en los modelos de datos / procesos.

Por estas razones, el proceso de obtención de los datos será el siguiente:

1. El administrador del sistema comprobará la publicación de los datos en junio de cada año.

2. El administrador del sistema comprobará que los datos tienen el mismo esquema que el año anterior. Esta comprobación será un híbrido entre la revisión de la guía de metodología / diccionario de datos [5] y los procesos de carga, que tendrán que verificar que el formato es el esperado.
3. El administrador lanzará los procesos de ETL que se definen en el apartado 3.3.

3.4. Diseño de modelo de datos

A partir de la información de la guía de metodología / diccionario de datos [5], se ha elaborado la tabla del Anexo 10 debido, que contiene la siguiente información:

- `Original field name`: nombre del campo dentro del fichero de Open Payments.
- `Original data type`: tipo de datos indicado por Open Payments.
- `B, C, D, E, F, G`: ficheros de datos en los que aparece cada campo codificados con el apéndice en el que se explican los diferentes archivos de datos de Open Payments:
    - Appendix B: General Payments Detail (Program Year 2016 and Upcoming Years)
    - Appendix C: General Payments Detail (Program Years 2013-2015)
    - Appendix D: Research Payments Detail (Program Year 2016 and Upcoming Years)
    - Appendix E: Research Payments Detail (Program Years 2013-2015)
    - Appendix F: Physician Ownership Information Detail (All Program Years)
    - Appendix G: Deleted and Removed Records File
- `DB`: indica los datos útiles para establecer relaciones entre las diferentes entidades. Estos datos se han duplicado, ya que deberán estar en varios ámbitos para actuar como claves primarias o foráneas.
- `Table`: tabla en la que se almacena el campo dentro de la base de datos del sistema de explotación desarrollado.
- `Field`: nombre del campo dentro de la base de datos del sistema de explotación desarrollado.
- `Data type`: tipo de datos dentro de la base de datos del sistema de explotación desarrollado.

Se han identificado mediante el símbolo (1) los datos relativos a una entidad que puede aparecer más de una vez para un pago. Por ejemplo, en un pago asociado a una investigación pueden indicarse hasta 5 investigadores, por lo que en el

fichero de Open Payments aparecen los campos `Principal_Investigator_X_Profile_ID`, donde X vale entre 1 y 5.

Por otro lado, se han identificado con el símbolo (2) aquellos datos que cambian de nomenclatura en 2016 pero que se considera que se pueden unificar:

- `Name_of_Associated_Covered_Drug_or_BiologicalX` y `Name_of_Associated_Covered_Device_or_Medical_SupplyX` se integran en `Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_X`.
- `NDC_of_Associated_Covered_Drug_or_BiologicalX` se integra en `Associated_Drug_or_Biological_NDC_X`.

Los datos del archivo suplementario relativo a los datos de los médicos no se utilizarán, ya que se consideran suficientes los datos incluidos en las transacciones para cumplir los objetivos establecidos.

A partir de la información de la tabla del Anexo 10 se elabora el modelo de datos representado en la Figura 2.

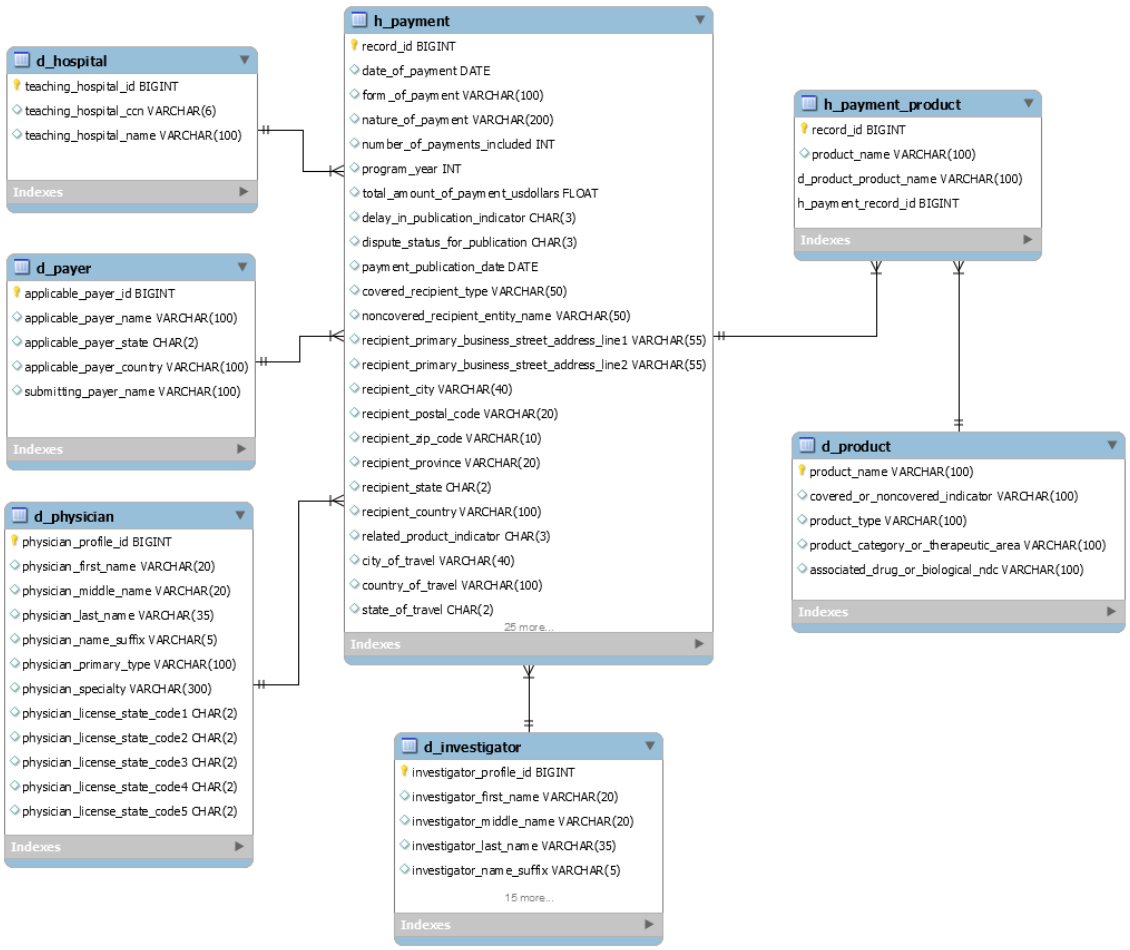


Figura 2 - Modelo de datos

Las tablas incluidas en el modelo de datos son:

- `h_payment`: cada registro representa un pago único.
- `h_payment_product`: almacena pares pago-producto. Cada registro representa un producto incluido en un pago.
- `d_product`: cada registro representa un producto único.
- `d_hospital`: cada registro representa un hospital único.
- `d_payer`: cada registro representa a un pagador único.
- `d_physician`: cada registro representa a un médico único.
- `d_investigator`: cada registro representa a un investigador único.

Las relaciones que se establecen entre las tablas son las siguientes:

- `h_payment` y `h_payment_product` se relacionan mediante `record_id`.
- `h_payment_product` y `d_product` se relacionan mediante `product_name`.
- `h_payment` y `d_hospital` se relacionan mediante `teaching_hospital_id`.
- `h_payment` y `d_payer` se relacionan mediante `applicable_payer_id`.
- `h_payment` y `d_physician` se relacionan mediante `physician_profile_id`.
- `h_payment` y `d_investigator` se relacionan mediante `investigator_profile_id`.

En el Anexo 11, se muestra el código SQL que implementa el modelo de datos en la base de datos. En él se pueden observar los campos que incluye cada tabla, el tipo de datos de cada campo y sus características y las relaciones entre tablas.

3.5. Diseño del sistema de almacenamiento

Para almacenar los datos se ha elegido una base de datos relacional, en concreto MySQL [6] por las siguientes razones:

- Es una herramienta open source.
- Es una herramienta consolidada en el mercado.
- Cuenta con una amplia comunidad de desarrolladores.

Por otro lado, se ha decidido generar dos esquemas de datos independientes:

- Un esquema para alojar los datos importados de Open Payments.
- Un esquema para alojar los indicadores calculados.

Esta separación en dos esquemas permitirá administrar de manera separada los permisos otorgados a los diferentes usuarios, dotando de mayor versatilidad y seguridad al sistema.

3.6. Diseño de los procesos de ETL

Como se ha comentado en el apartado anterior, el administrador del sistema comprobará la publicación de los datos en junio de cada año y que los archivos mantienen la estructura esperada. Por lo tanto, no ha sido necesario diseñar ningún sistema de ingesta específico ya que la ingesta será manual.

Para el proceso de ETL se ha elegido Talend Open Studio for Data Integration [7] por las siguientes razones:

- Es una herramienta open source.
- Es una herramienta en constante desarrollo y adaptada a las últimas tecnologías de bases de datos / datawarehouse.
- Cuenta con una amplia comunidad de desarrolladores.
- Es un software de alto nivel, lo que permitiría a futuros desarrolladores incorporarse rápidamente al proyecto.

Conceptualmente, el proceso de ETL utilizará como entrada los ficheros CSV publicados por Open Payments y escribirá los datos procesados directamente en la base de datos que los almacenará.

3.7. Diseño del sistema de análisis

El sistema de análisis no requiere de ningún módulo adicional, ya que el análisis se realizará mediante las utilidades de consulta de la propia base de datos MySQL y los indicadores calculados se almacenarán en la propia base de datos.

3.8. Diseño de la visualización de los datos analizados

Para visualizar los datos analizados se ha elegido la plataforma de inteligencia de negocio Qlik Sense [8] por las siguientes razones:

- Es una herramienta gratuita sin limitación de funcionalidad en su versión escritorio.
- Es una herramienta consolidada en el mercado.
- Permite conexión con MySQL mediante un conector ODBC.
- La versión gratuita de la Qlik Sense Cloud es potente y ofrece muchas funcionalidades.
- No solo permite visualización, sino que permite realizar analítica de los datos.

- La versión de pago de Qlik Sense Cloud tiene un coste asumible en caso de necesitar extender el sistema de analítica a más de 5 usuarios.
- Cuenta con una amplia comunidad de desarrolladores.

A pesar de que Qlik Sense tiene capacidad analítica, se ha decidido procesar los datos directamente en MySQL debido al gran volumen de datos manejados. Sin embargo, se pretende que algunos indicadores se calculen relativamente desagregados para que sea Qlik Sense el que termine de agregarlos de la manera que requiera el usuario que analiza la información.

4. Implementación del sistema

Como el tiempo disponible para la elaboración del presente trabajo es muy limitado, se ha optado por implementar la arquitectura en local. Las características de la máquina utilizada son las siguientes:

- Procesador Intel Core i5-7200U (2 núcleos, 3MB cache, 2.5GHz hasta 3.1GHz) [9]
- Memoria RAM 8GB DDR4 2133MHz
- Disco duro 256GB SSD M.2 SATA3
- Sistema operativo Windows 10 Home.

4.1. Implementación del sistema de almacenamiento

Para la implementación del sistema de almacenamiento se ha utilizado MySQL Community Server versión 8.0.13 (la más reciente en el momento de la implementación) [10].

Para la configuración del servidor MySQL y de la base de datos se ha utilizado MySQL Workbench versión 8.0.13 (la más reciente en el momento de la implementación) [11].

Mediante MySQL Workbench se han llevado a cabo las siguientes tareas:

- Definir todas las tablas previstas en el modelo de datos:
  - Definir los campos tal cual se detallan en el Anexo 10.
  - Definir las relaciones entre tablas descritas en el apartado 3.4.
- Crear un script que implemente:
  - El modelo de datos en la base de datos sobre el esquema `open_payments`, que contendrá los datos importados de Open Payments.
  - El esquema `indicators`, que contendrá la tabla con los indicadores calculados.

  Este script se incluye en el Anexo 11.

- Crear los siguientes usuarios para la base de datos:
    - `admin`: usuario con privilegios completos, tanto de administración de la base de datos como de derechos de usuarios en todos los esquemas.
    - `talend`: usuario con privilegios de inserción, actualización, borrado y consulta de datos en el esquema `open_payments`. Este usuario es el que se utilizará en los procesos de ETL.
    - `analytics`: usuario con privilegios de consulta a todos los datos del esquema `open_payments` y con privilegios completos en el esquema `indicators`. Este usuario es el que se utilizará para analizar los datos.
    - `viewer`: usuario con privilegios de consulta para todos los datos del esquema `indicators`. Este usuario es el que utilizará el sistema de visualización de los indicadores calculados.
- Crear el esquema de base de datos `open_payments` mediante el script del Anexo 11. Una vez ejecutado este punto, ya se dispone de un entorno útil para insertar datos.
- Ajustar el timeout de lectura de la base de datos [12]. Por defecto, MySQL lo define a 30 segundos. Sin embargo, como la base de datos es grande, algunas consultas después de realizar la carga inicial requieren más tiempo. Se configura de modo que no se produzca nunca timeout de lectura (`DBMS connection read timeout Interval = 0`).
- Crear una copia de seguridad una vez realizada la carga inicial [13].

Después de cargar todos los datos tal y como se explica en los apartados posteriores, el tamaño de la base de datos en disco es de 135GB.

4.2. Implementación de los procesos de ETL

Para la implementación de los procesos de ETL se ha utilizado Talend Open Studio for Data Integration versión 7.1.1 (la más reciente en el momento de la implementación) [7].

En primer lugar, se han definido los formatos de las conexiones con la base de datos y de los ficheros CSV. En la conexión con la base de datos se gestionan todas las tablas. En cambio, cada tipo de fichero CSV tiene una descripción de metadatos diferente. Si un tipo de fichero que tiene diferente formato en función del año, tiene múltiples descripciones diferentes. En el Anexo 12, se adjuntan los ficheros XML que describen estos formatos.

A continuación, se han creado todos los subtrabajos necesarios para realizar la carga inicial de los datos. En el Anexo 13 se adjuntan todos los subtrabajos creados. Todos los subtrabajos tienen una estructura similar. La Figura 3 utiliza un extracto de un subtrabajo para ilustrar dicha estructura.

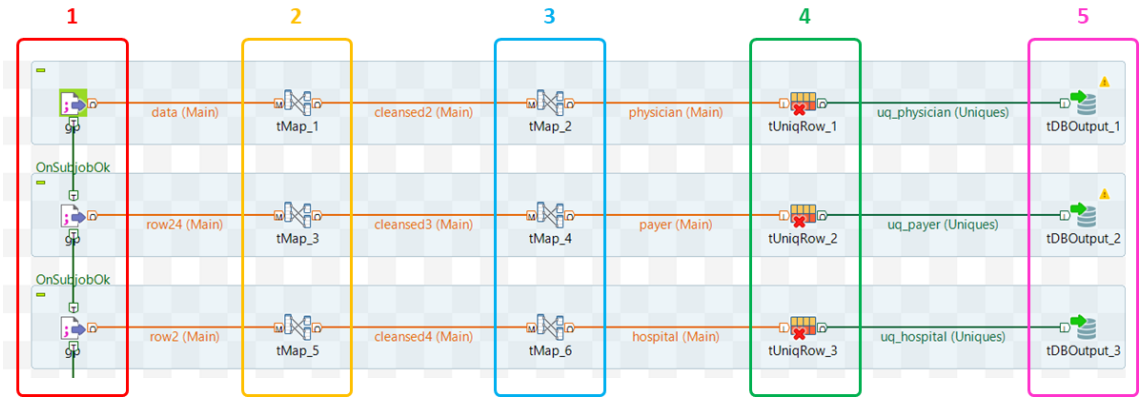


Figura 3 - Ejemplo de estructura de subtrabajos de la ETL

Los diferentes bloques realizan las siguientes tareas:

1. Lectura de los ficheros CSV mediante objetos `tFileInputDelimited`.
2. Procesado de los datos contenidos en el fichero CSV mediante objetos `tMap`.
3. Extracción de los datos relativos a una entidad concreta (hospital, médico, investigador, pago, etc.) mediante objetos `tMap`.
4. Eliminación de duplicados (se realiza para todas las entidades excepto para los pagos) mediante objetos `tUniqRow`.
5. Escritura en base de datos mediante objetos `tDBOutput`.

El trabajo ETL que realiza el proceso completo de carga inicial de datos es el mostrado en el Anexo 13.20.

La ejecución del proceso completo de carga abortaba pasadas unas 24 horas debido a problemas de memoria. Debido a esto, se ha decidido fragmentar los archivos más grandes, en concreto los de pagos generales, en varios archivos de menor tamaño para evitar el problema. Para ello, se ha diseñado un script en Python, adjunto en el Anexo 14.

Para ahorrar tiempo en el proceso de carga, una vez realizada la fragmentación de los archivos no se han rehecho los trabajos de Talend, si no que se han aprovechado los subtrabajos existentes ejecutándolos de manera manual (renombrando los fragmentos con los nombres del fichero completo y ejecutando más de una vez).

Por otro lado, se ha observado que el procesado de los ficheros CSV en Talend mediante los objetos `tFileInputDelimited` tiene alguna inconsistencia:

- En campos de texto delimitados entre dobles comillas que incluyen comas en el texto:

  ```
  ..., "blablabla, blablabla", ...
  ```

- En campos de texto delimitados entre dobles comillas que incluyen dobles comillas en el texto:

  ```
  ..., "blablabla "blablabla" blablabla", ...
  ```

Se ha probado a sustituir objetos `tFileInputDelimited` por la secuencia de objetos `tFileInputFullRow`, `tReplace` y `tExtractDelimitedFields`. El objetivo ha sido leer el fichero CSV línea a línea, sustituir mediante expresiones regulares las comas y comillas que daban problemas y luego realizar el procesado habitual. La sustitución con expresiones regulares realizada ha sido:

- `"([^\",])(,)([^\",])"` → `"$1;$3"`
- `"[\"]"` → `""`

El resultado de esta prueba no ha sido satisfactorio, obteniendo un número de errores mayor que el obtenido utilizando el fichero original, por lo que, dado que el número de errores es muy bajo, se decidió asumir la pérdida de estos pagos. Además, los tiempos de carga con esta alternativa eran sensiblemente superiores.

En el Anexo 15, se adjuntan una tabla con los tiempos de carga de los diferentes trabajos de la carga inicial. En resumen, el proceso de carga:

- Ha tardado 4 días, 13 horas y 27 minutos en procesar y guardar los datos en la base de datos (unos 7 días de trabajo debido a la actividad manual).
- Ha procesado 53,013,925 pagos.
- Ha encontrado 128 incoherencias de formato en registros de pago, con lo que estos pagos se han desechado.

Para los datos futuros, en caso de que no haya cambios de formatos, se actualizarán los ficheros CSV de entrada en los trabajos indicados a continuación para generar los trabajos de los años siguientes:

- `j16_delrem_2016` → `j20_delrem_2017`
- `j17_gp_f16_2017` → `j21_gp_f16_2018`
- `j18_gp_r16_2017` → `j22_rp_f16_2018`
- `j19_poi_2017` → `j22_poi_2018`

4.3. Implementación del sistema de análisis

En la fase de diseño se definió un modelo de datos en el que la mayor parte de la información relativa a pagos se concentraba en la tabla `h_payment`. Sin embargo, debido a las limitaciones de equipamiento del presente trabajo (tener que usar un ordenador personal) ha resultado inviable realizar consultas directamente sobre esta tabla.

Para solventar este problema, se ha fragmentado el contenido de `h_payment` en tablas más pequeñas que contienen la información necesaria para análisis concretos y divididas con los datos de un único año. Dichas tablas son:

- `h_payment_details_XXXX`: para cada pago, se incluye quien lo realiza, su importe y los datos relativos al pago (naturaleza, forma, etc.).
- `h_payment_physician_XXXX`: para cada pago a un médico, se incluye quien lo realiza, su importe, el médico que lo recibe, su estado y su país.
- `h_payment_hospital_XXXX`: para cada pago a un hospital, se incluye quien lo realiza, su importe, el hospital que lo recibe, su estado y su país.
- `h_payment_noncovered_XXXX`: para cada pago a un receptor no cubierto, se incluye quien lo realiza, su importe, el receptor, su estado y su país.
- `h_payment_investigator_XXXX`: para cada pago en el que interviene un investigador, se incluye quien lo realiza, su importe, el médico u hospital que recibe el pago, su estado y su país y el investigador implicado.
- `h_payment_physician_3p_XXXX`: para cada pago a terceros a través de un médico, se incluye quien lo realiza, su importe, el médico que lo recibe, su estado y su país y los datos del tercero.
- `h_payment_hospital_3p_XXXX`: para cada pago a terceros a través de un hospital, se incluye quien lo realiza, su importe, el hospital que lo recibe, su estado y su país y los datos del tercero.
- `h_payment_poi_XXXX`: para cada pago a un médico con intereses en el pagador, se incluye quien lo realiza, su importe, el médico implicado, el estado y el país del receptor y los datos relativos al interés.
- `h_payment_travel_XXXX`: para cada pago asociado a un viaje, se incluye quien lo realiza, su importe y los datos relativos al viaje.
- `h_payment_list_physician_XXXX`: para cada pago a un médico, se incluye quien lo realiza, su importe y y el estado y país del médico.

`XXXX` representa el año de los datos incluidos (de 2013 a 2017). La creación de estas tablas está incluida en el script del Anexo 11. El script que rellena los datos de estas tablas a partir del contenido de `h_payment` se adjunta en el Anexo 16.

Es importante destacar que la gestión de estas tablas adicionales se ha desvinculado de la ETL para dar flexibilidad al analista para realizar los cambios que necesite durante el proceso de análisis. Esta desvinculación de la ETL provoca que, cuando haya una nueva publicación de datos, además de la ETL, se tenga que ejecutar el script de carga para el nuevo año.

4.4. Implementación de la visualización de los datos analizados

Para la visualización de los datos se ha instalado Qlik Sense [14]. Posteriormente, se ha instalado y configurado el conector ODBC que permite a Qlik Sense acceder a los datos disponibles en MySQL [15] [16].

El acceso de Qlik Sense a la base de datos se hace mediante el usuario `viewer`. De este modo se consigue:

- Evitar sobrecargas en la base de datos, ya que el usuario `viewer` solo tiene acceso al esquema `indicators` (con poco volumen de datos), no al esquema `open_payments` (que tiene un gran volumen de datos).
- Proteger los indicadores calculados, ya que el usuario `viewer` únicamente tiene permisos de consulta.

La estrategia de visualización de la información generada a partir de los datos de Open Payments ha consistido en la creación de una aplicación que incluye diferentes tableros, uno por pregunta analítica a responder. Se ha establecido esta limitación de un tablero por pregunta para ser concisos y que los indicadores mostrados sean relevantes.

5. Respuesta a las preguntas analíticas

5.1. ¿Qué tipos de pagos son los más repetidos?

Para cada año, tipo de beneficiario, forma de pago y naturaleza del pago, se ha calculado el número de pagos realizados y suma del importe de los mismos. El resultado se ha guardado en la tabla `indicators.type_analysis`. El script que realiza estos cálculos se adjunta en el Anexo 17.1.

El tablero para el análisis por tipo de pagos, incluido en el Anexo 18.1, contiene los siguientes gráficos:

- Número de pagos recibidos (absoluto y relativo) por tipo de receptor y año.
- Importe de los pagos recibidos (absoluto y relativo) por tipo de receptor y año.
- Importe promedio de los pagos recibidos por tipo de receptor y año.
- Distribución del número de pagos recibidos por forma de pago y año.
- Distribución del número de pagos recibidos por naturaleza del pago y año.

Este tablero permite filtrar por año y por tipo de receptor (`H`: hospital, `N`: receptor no cubierto por el programa, `P`: médico).

En la Figura 4 se muestra que la mayoría de pagos se realiza a médicos, mientras que aproximadamente la mitad del importe de los pagos se destina a entidades no cubiertas.

Desde el punto de vista temporal, se observa que, en los últimos años, tanto entidades no cubiertas como médicos han recibido un menor número de pagos y menos importe que en años anteriores. En cambio, el número de pagos a hospitales ha crecido ligeramente.

Esta figura también muestra que el pago promedio a los hospitales es el más elevado, aproximadamente el doble del pago medio a entidades no cubiertas y unas 50 veces mayor que el pago a médicos. También se observa una tendencia creciente para hospitales y entidades no cubiertas, mientras que para médicos se mantiene estable.

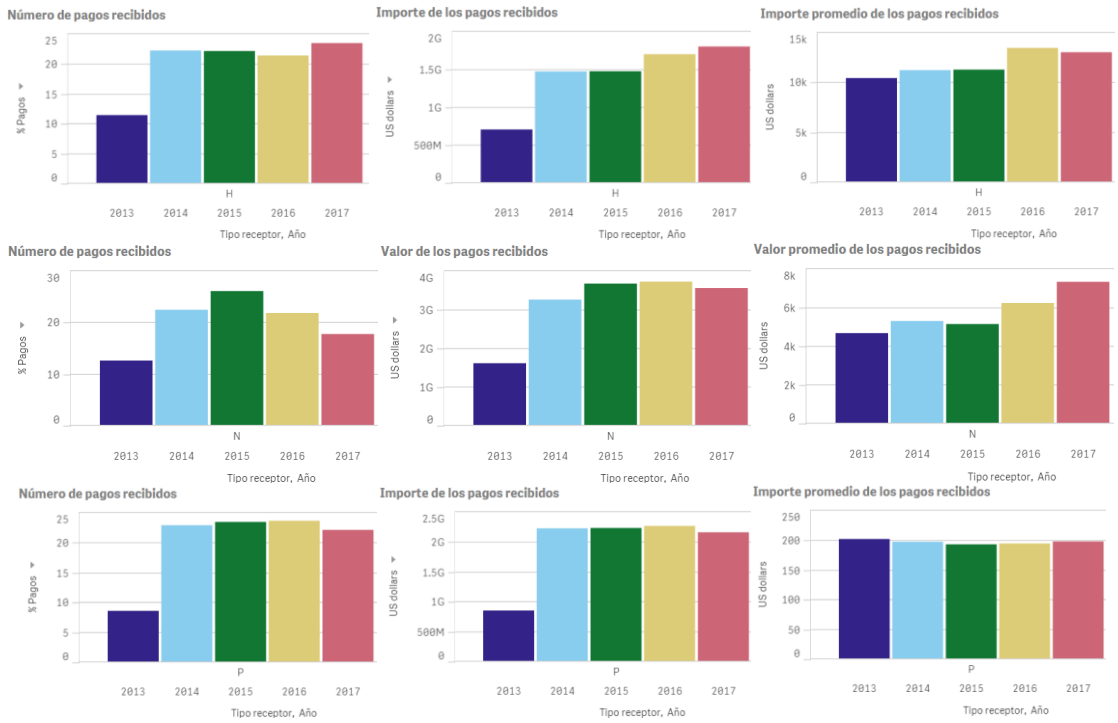


Figura 4 - Número de pagos, importe total e importe medio por pago en función del tipo de receptor y del año

Por último, en la Figura 5 se observa cómo las distribuciones de la forma de pago y de la naturaleza de los pagos son similares para los diferentes años. En la Figura 6 se aprecia que:

- Las formas de pago son fundamentalmente en especie y en metálico.
- La mayor parte de los pagos son de naturaleza “comida y bebida” y “viajes y alojamiento”.

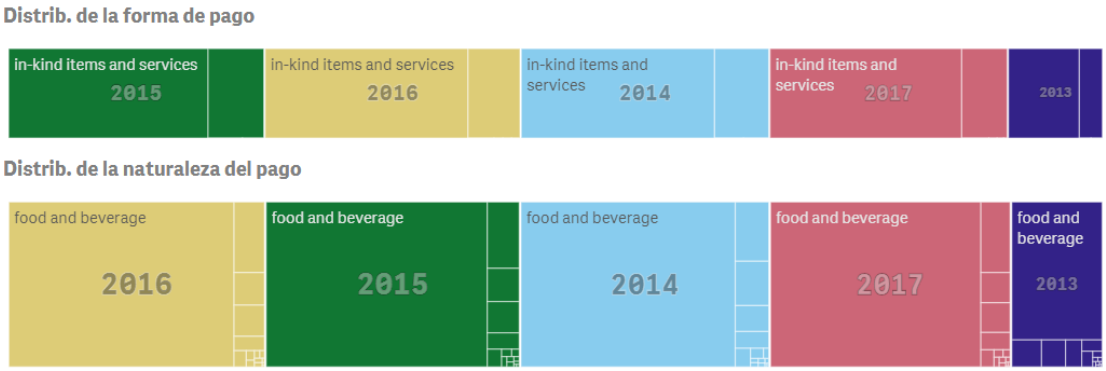


Figura 5 - Distribución de la forma y la naturaleza de los pagos por año

Figura 6 - Distribución de la forma y la naturaleza de los pagos en 2017

5.2. ¿Qué países y estados son los más influenciados?

Para cada año, tipo de beneficiario y estado y país del receptor del pago, se ha calculado el número de pagos realizados y la suma del importe de los mismos. El resultado se ha guardado en la tabla `indicators.location_analysis`. El script que realiza estos cálculos se adjunta en el Anexo 17.2.

El tablero para el análisis por localización, incluido en el Anexo 18.2, contiene los siguientes elementos:

- Una tabla que muestra el número de pagos recibidos (absoluto y relativo), la suma del importe de todos los pagos (absoluto y relativo) y el importe medio por pago por país.
- Un mapa con la distribución de la suma del importe de todos los pagos recibidos por país.
- Un mapa con la distribución de la suma del importe de todos los pagos recibidos por estado de Estados Unidos.

Este tablero permite filtrar por año, por tipo de receptor (`H`: hospital, `N`: receptor no cubierto por el programa, `P`: médico), por país y por estado (sólo para Estados Unidos).

La información mostrada en la Figura 7 expone que Estados Unidos es el receptor de la mayoría de pagos. Esto sucede para todos los tipos de receptores y para todos los años.

Figura 7 - Distribución del importe de los pagos entre 2013 y 2017

El resto de países receptores de pagos están distribuidos por todo el mundo. En la Tabla 1 se muestra que, si se obvia Estados Unidos, los países que han recibido el mayor número de pagos son Reino Unido, las islas periféricas menores de Estados Unidos (no incorporadas a los Estados Unidos [17]) y Canadá. En cambio, si se observa desde el punto de vista del importe de los pagos, como se presenta en la Tabla 2, los países que han recibido mayor importe han sido Canadá, Reino Unido y Alemania.

País	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales	**3,683.00**	**100.00**	**7,750,482.56**	**100.00**	**2,104.39**
united kingdom	1,281.00	34.78	681,090.17	8.79	531.69
united states minor outlying islands	856.00	23.24	162,976.58	2.10	190.39
canada	707.00	19.20	5,278,363.39	68.10	7,465.86

Tabla 1 - Top 3 de los países que reciben más pagos

País	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales	**3,683.00**	**100.00**	**7,750,482.56**	**100.00**	**2,104.39**
canada	707.00	19.20	5,278,363.39	68.10	7,465.86
united kingdom	1,281.00	34.78	681,090.17	8.79	531.69
germany	127.00	3.45	577,703.65	7.45	4,548.85

Tabla 2 - Top 3 de los países que reciben mayores importes

En la Figura 8, se observa que, sin distinguir el tipo de receptor, los países beneficiarios se distribuyen por todos los continentes. En el caso de que los receptores sean médicos, se mantiene una distribución similar. En cambio, los pagos a entidades no cubiertas se concentran en Canadá y Europa.

En el caso de pagos a hospitales, sólo se realizan a Ucrania y a las islas periféricas menores de Estados Unidos. Al haberse realizado únicamente dos pagos a hospitales fuera de Estados Unidos, no se considera muestra de una tendencia.

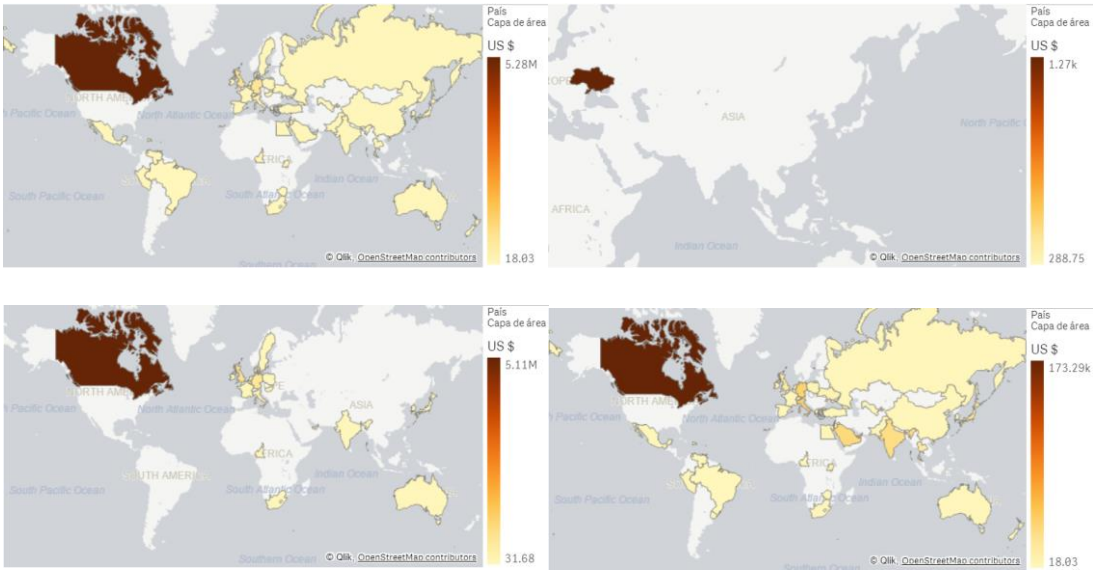


Figura 8 - Distribución del importe de los pagos fuera de EEUU por país (de izquierda a derecha y de arriba abajo, total, hospitales, entidades no cubiertas y médicos)

En la Figura 9 y la Figura 10, se observa que, en Estados Unidos, a nivel global, los estados más beneficiados por el programa son California, Texas, Florida y Nueva York. Esta distribución se mantiene cuando los receptores de los pagos son médicos o entidades no cubiertas. En cambio, cuando los receptores de los pagos son hospitales, los estados más beneficiados por el programa son California, Texas y Massachusetts.

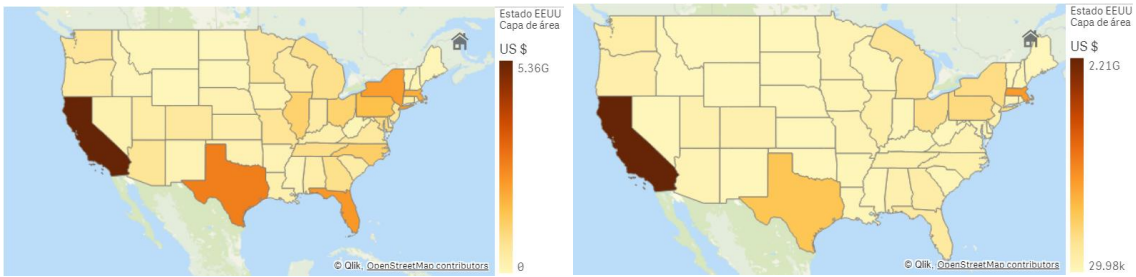


Figura 9 - Distribución del importe de los pagos en EEUU por estados (de izquierda a derecha, total y hospitales)

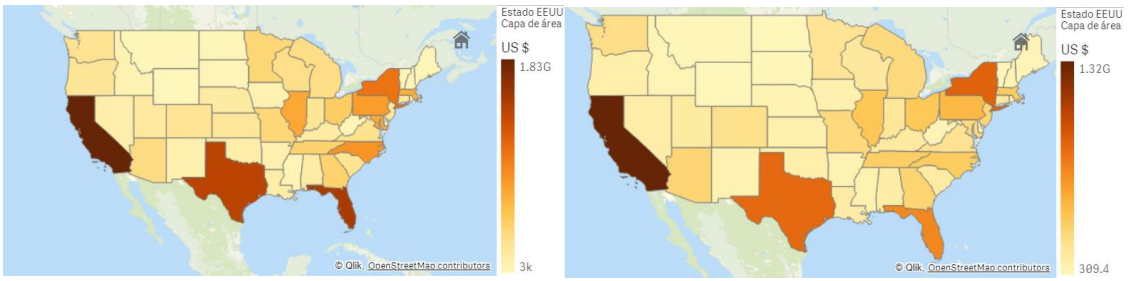


Figura 10 - Distribución del importe de los pagos en EEUU por estados (de izquierda a derecha, entidades no cubiertas y médicos)

5.3. ¿Qué receptores e investigadores son los más influenciados?

Para cada año, tipo de entidad y estado y país de la entidad, se ha calculado el número de pagos realizados y suma del importe de los mismos. El resultado se ha guardado en la tabla `indicators.ranking_analysis`. El script que realiza estos cálculos se adjunta en el Anexo 17.3.

El tablero que muestra el ranking por entidad, incluido en el Anexo 18.3, contiene una tabla que muestra:

- La entidad y el tipo al que pertenece.
- El país y el estado (si es de Estados Unidos) al que pertenece la entidad.
- El número de pagos (absoluto y relativo) recibidos por la entidad en los años seleccionados.
- La suma del importe de todos los pagos (absoluto y relativo) recibidos por la entidad en los años seleccionados.
- El importe medio por pago a la entidad en los años seleccionados.

Es importante explicar que se han incluido en la misma tabla hospitales, médicos, entidades no cubiertas e investigadores. Los investigadores no reciben pagos. Por lo tanto, al realizar análisis por investigador lo que se muestra es el número e importe de los pagos en los que interviene un investigador. De este modo, los totales obtenidos teniendo en cuenta todos los tipos de entidad no son los reales. (sí que lo son si se excluye a los investigadores).

Este tablero permite filtrar por año, por tipo de receptor (`H`: hospital, `N`: receptor no cubierto por el programa, `P`: médico, `I`: investigador), por país y por estado (sólo para Estados Unidos).

En las siguientes tablas se muestra el top 5 de hospitales, médicos y entidades no cubiertas que más importe han recibido entre 2013 y 2017. Además, también se presenta el top 5 de investigadores asociados a mayor importe de pagos en el mismo periodo.

Receptor	Tipo	País	Estado	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales				**593,497**	**100.00**	**7,100,646,283**	**100.00**	**11,964**
city of hope national medical center	H	united states	California	2,768	0.47	1,051,459,051	14.81	379,862
ut md anderson cancer center	H	united states	Texas	12,072	2.03	440,933,652	6.21	36,525
city of hope national medical cnt	H	united states	California	1,198	0.20	419,629,412	5.91	350,275
hospital of the univ of penna	H	united states	Pennsylvania	10,327	1.74	267,865,375	3.77	25,938
massachusetts general hospital	H	united states	Massachusetts	7,006	1.18	264,044,791	3.72	37,688

Tabla 3 - Top 5 de los hospitales que reciben mayor importe en pagos

Receptor	Tipo	País	Estado	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales				**2,758,519**	**100.00**	**15,714,180,301**	**100.00**	**5,697**
duke university	N	united states	North Carolina	6,212	0.23	140,695,895	0.90	22,649
national cancer institute	N	united states	Maryland	139	0.01	130,124,455	0.83	936,147
dana farber cancer institute	N	united states	Massachusetts	3,649	0.13	127,999,867	0.81	35,078
memorial sloan kettering cancer center	N	united states	New York	4,472	0.16	105,343,433	0.67	23,556
swog cti	N	united states	Michigan	10	0.00	90,799,274	0.58	9,079,927

Tabla 4 - Top 5 de las entidades no cubiertas que reciben mayor importe en pagos

Receptor	Tipo	País	Estado	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales				**49,637,693**	**100.00**	**9,669,622,025**	**100.00**	**195**
burkhart, stephen (288926)	P	united states	Texas	281	0.00	88,224,493	0.91	313,966
foley, kevin (311622)	P	united states	Tennessee	294	0.00	76,827,270	0.79	261,317
jackson, roger (354917)	P	united states	Missouri	146	0.00	67,451,922	0.70	461,999
narayan, sujata (281659)	P	united states	California	13	0.00	57,717,989	0.60	4,439,845
underwood, karen (933844)	P	united states	Arizona	12	0.00	37,558,631	0.39	3,129,886

Tabla 5 - Top 5 de los médicos que reciben mayor importe en pagos

Receptor	Tipo	País	Estado	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales				**3,164,837**	**100.00**	**19,632,022,110**	**100.00**	**6,203**
ryan, christopher (183323)	I	united states	Oregon	421	0.01	147,970,092	0.75	351,473
doroshow, james (826196)	I	united states	Maryland	20	0.00	113,391,671	0.58	5,669,584
cortes-franco, jorge (26501)	I	united states	Texas	641	0.02	98,953,212	0.50	154,373
bertagnolli, monica (114497)	I	united states	Massachusetts	197	0.01	87,383,008	0.45	443,569
patel, manesh (340070)	I	united states	North Carolina	132	0.00	76,969,590	0.39	583,103

Tabla 6 - Top 5 de los investigadores implicados en mayor importe

La hoja se ha diseñado para permitir la mayor flexibilidad posible, por ejemplo, hacer este análisis para un año concreto, para un país o estado concretos, etc.

5.4. ¿Cuáles son las terceras partes más beneficiadas de los pagos? ¿Tienen las aportaciones a caridad un peso importante en el programa?

Para cada año, tipo de beneficiario, estado y país del receptor del pago, tercera parte e indicador de si es una aportación a la caridad, se ha calculado el número de pagos realizados y suma del importe de los mismos. El resultado se ha guardado en la tabla `indicators.third_party_analysis`. El script que realiza estos cálculos se adjunta en el Anexo17.4.

El tablero para el análisis de terceras partes, incluido en el Anexo 18.4, contiene los siguientes elementos:

- Un gráfico de barras que muestra la suma del importe de todos los pagos realizados a terceras partes en función de si son aportaciones a la caridad o no.

- Un gráfico circular que muestra el porcentaje que representa el importe de los pagos a terceras partes en función de si son aportaciones a la caridad o no.
- Un mapa con la distribución de la suma del importe de todos los pagos realizados a terceras partes correspondientes a aportaciones a la caridad por estado de Estados Unidos.
- Una tabla que nuestra:
    - El nombre de la tercera parte y el tipo al que pertenece la entidad que recibe el pago para transferirlo a la tercera parte.
    - El país y el estado (si es de Estados Unidos) al que pertenece la entidad que recibe el pago para transferirlo a la tercera parte.
    - El número de pagos (absoluto y relativo) realizados a la entidad que recibe el pago en los años seleccionados.
    - La suma del importe de todos los pagos (absoluto y relativo) realizados a la entidad que recibe el pago en los años seleccionados.
    - El importe medio por pago a la entidad que recibe el pago en los años seleccionados.

Este tablero permite filtrar por año, por tipo de receptor (`H`: hospital, `N`: receptor no cubierto por el programa, `P`: médico), por país y por estado (sólo para Estados Unidos).

La información mostrada en la Figura 11 evidencia que:

- El importe de los pagos a terceras partes a través de los médicos es mucho mayor que el de los pagos a través de los hospitales.
- El porcentaje de aportaciones a caridad de mediante pagos a terceras partes a través de hospitales es mayor que el realizado a través de médicos. De todos modos, el importe de las contribuciones a caridad a través de médicos es mayor que a través de hospitales.

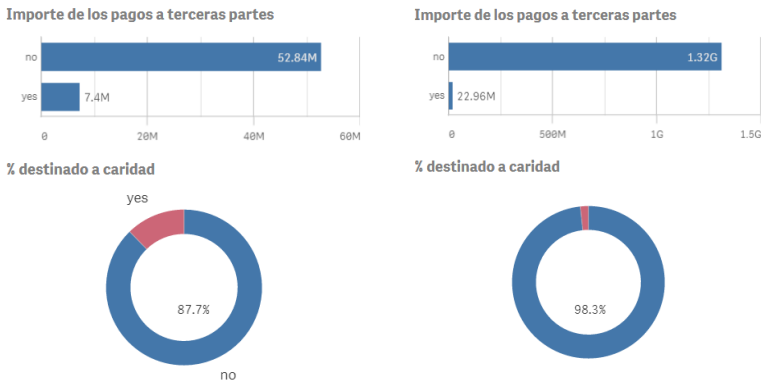


Figura 11 - Importe y proporción de los pagos a terceras partes destinado a la caridad (a la izquierda, hospitales; a la derecha, médicos)

En la Figura 12 se observa la tendencia creciente de la proporción de las aportaciones a caridad durante el periodo 2013-2016, aunque en 2017 cayeron a proporciones del 2014.

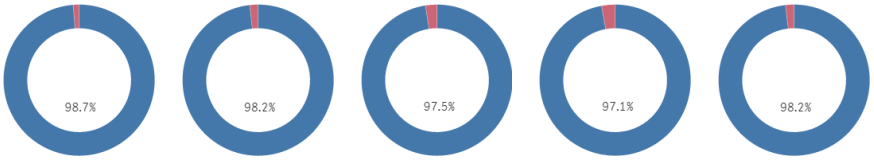


Figura 12 - Proporción de los pagos a terceras partes destinado a la caridad (de izquierda a derecha, 2013, 2014, 2015, 2016 y 2017)

En la Figura 13 se observa que, en Estados Unidos, la distribución de las aportaciones a caridad por estado varía cada año. A pesar de ello, los estados de California, Florida y Nueva York suelen destacar.

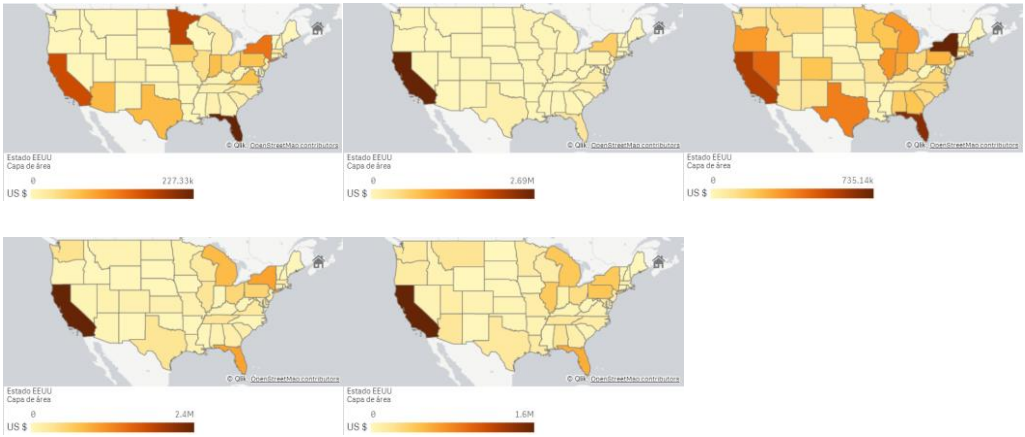


Figura 13 - Distribución de las aportaciones a caridad en EEUU por estados (de izquierda a derecha y de arriba abajo, 2013, 2014, 2015, 2016 y 2017)

Por último, en las siguientes tablas se muestra el top 5 de terceras partes que más importe han recibido entre 2013 y 2017 distinguiendo si son aportaciones a caridad o no lo son.

Tercera parte	Receptor	País	Estado	Caridad	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales					**615,942**	**100.00**	**1,368,916,349**	**100.00**	**2,222**
burkhart resource, ltd	P	united states	Texas	no	62	0.01	87,761,169	6.41	1,415,503
guidance endodontics llc	P	united states	New Mexico	no	1	0.00	22,880,194	1.67	22,880,194
mayo foundation	P	united states	Minnesota	no	256	0.04	22,646,567	1.65	88,463
hackensack university medical center foundation	H	united states	New Jersey	no	2	0.00	15,004,000	1.10	7,502,000
james g berbee and karen a walsh joint revocable t	P	united states	Wisconsin	no	3	0.00	14,265,661	1.04	4,755,220

Tabla 7 - Top 5 de las terceras partes que han recibido mayor importe no asociado a caridad

Tercera parte	Receptor	País	Estado	Caridad	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales					**3,174**	**100.00**	**30,358,150**	**100.00**	**9,565**
the sujata and sanjiv narayan foundation	P	united states	California	yes	1	0.03	2,445,626	8.06	2,445,626
annenberg center for health sciences at eisenhower	H	united states	California	yes	5	0.16	1,289,928	4.25	257,986
sujata and sanjiv narayan foundation	P	united states	California	yes	2	0.06	736,195	2.43	368,098
kenya relief	P	united states	Michigan	yes	5	0.16	727,976	2.40	145,595
stony brook foundation, inc.	H	united states	New York	yes	21	0.66	584,237	1.92	27,821

Tabla 8 - Top 5 de las terceras partes que han recibido mayor importe asociado a caridad

5.5. ¿Cuáles son los principales destinos de viajes pagados por el programa?

Para cada año, tipo de beneficiario y destino de viaje (ciudad, estado y país), se ha calculado el número de pagos realizados y suma del importe de los mismos. El resultado se ha guardado en la tabla `indicators.travel_analysis`. El script que realiza estos cálculos se adjunta en el Anexo 17.5.

El tablero para el análisis de los viajes, incluido en el Anexo 18.5, contiene los siguientes elementos:

- Una tabla que muestra el número de pagos asociados a viajes (absoluto y relativo), la suma del importe de todos los pagos asociados a viajes (absoluto y relativo) y el importe medio por pago asociado a viaje por ciudad.
- Un mapa con la distribución de la suma del importe de todos los pagos asociados a viajes por país.
- Un mapa con la distribución de la suma del importe de todos los pagos asociados a viajes por estado de Estados Unidos.
- Un mapa con la distribución de la suma del importe de todos los pagos asociados a viajes por ciudad.

Este tablero permite filtrar por año, por tipo de receptor (`H`: hospital, `N`: receptor no cubierto por el programa, `P`: médico) y por país.

En la Figura 14, se observa que la mayoría del importe de pagos asociados a viajes se dedica a viajes con destino a Estados Unidos. Además, los principales destinos son ciudades de los estados de California, Texas y Florida.

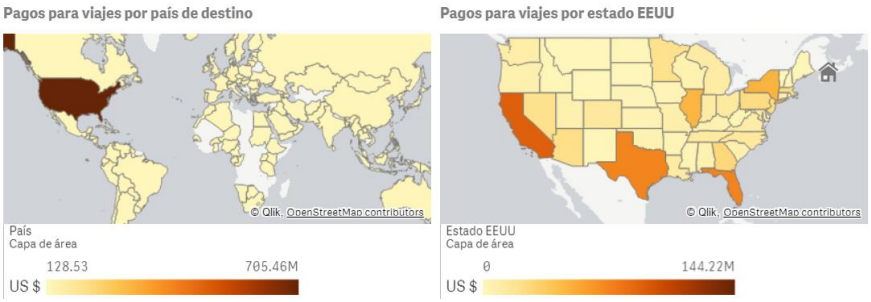


Figura 14 - Distribución de los pagos asociados a viajes por país (izquierda) y por estado de EEUU (derecha)

En la Figura 15, se aprecia que los principales destinos fuera de Estados Unidos son Alemania, Reino Unido, Japón, Francia y España.

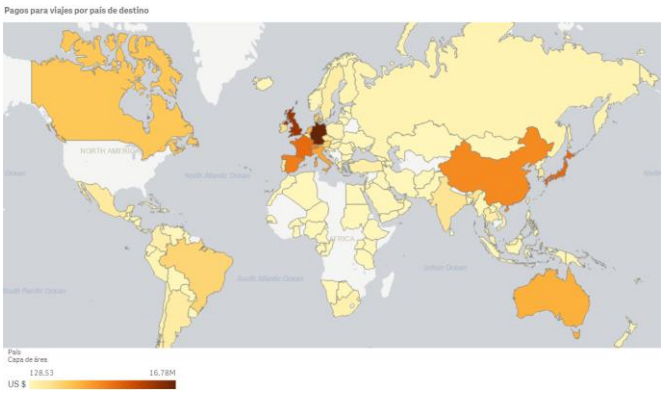


Figura 15 - Distribución de los pagos asociados a viajes fuera de Estados Unidos por país

En la Figura 16, se muestran las ciudades a las que se producen los viajes. En la Tabla 10, se listan las diez ciudades a las que se ha dedicado mayor importe en pagos.

Figura 16 - Distribución de los pagos asociados a viajes por ciudad

País	Estado	Ciudad	Pagos	US $	US $ / Pago
Totales			**2,522,150**	**849,670,422**	**337**
united states	Illinois	chicago	116,824	33,692,792	288
united states	New York	new york	84,993	29,939,204	352
united states	Texas	dallas	100,431	27,294,688	272
united states	Georgia	atlanta	77,884	20,732,031	266
united states	Nevada	las vegas	57,199	18,207,280	318
united states	Florida	miami	55,773	17,131,240	307
united states	California	san diego	54,600	16,210,661	297
united states	Florida	orlando	56,517	15,040,814	266
united states	Massachusetts	boston	45,757	14,888,647	325
united states	California	san francisco	41,437	14,588,122	352

Tabla 9 - Top 10 de las ciudades con mayor importe dedicado a viajes

5.6. ¿Qué relación tienen los intereses de los médicos en los pagadores y los pagos recibidos?

Para cada año, médico, pagador y estado y país del mismo, se ha calculado:

- Para cada par médico-pagador en el que el médico no tiene interés en el pagador, el importe total recibido.

- Para cada par médico-pagador en el que el médico tiene interés en el pagador, el importe total recibido, la cantidad que el médico tiene invertida en el pagador y la relación importe pagado sobre cantidad invertida:

  ```
  p_i_ratio = sum_payment / sum_invested
  ```

El resultado se ha guardado en la tabla `indicators.interest_analysis`. El script que realiza estos cálculos se adjunta en el Anexo 17.6.

El tablero para el análisis de los intereses de los médicos, incluido en el Anexo 18.6, incluye los siguientes elementos:

- Un indicador numérico que muestra el promedio del importe que recibe un médico por parte pagadores sobre los que tiene intereses.
- Un indicador numérico que muestra el promedio de la cantidad invertida por un médico en pagadores sobre los que tiene intereses.
- Un indicador numérico que muestra el promedio del importe que recibe un médico por parte de pagadores sobre los que no tiene intereses.
- Un indicador numérico que muestra el promedio de la relación importe recibido sobre la cantidad invertida para el caso de médicos por parte de pagadores sobre los que tiene intereses.
- Un gráfico de dispersión que muestra, para cada año, el promedio de los importes recibidos sobre las cantidades invertidas para el colectivo de médicos que tienen intereses en los pagadores.

Se ha considerado que un médico tiene intereses sobre un pagador si la cantidad invertida en el pagador es superior a un dólar estadounidense.

Observando los indicadores numéricos de la Figura 17, se concluye que:

- En media, los médicos que tienen intereses en los pagadores reciben importes muy superiores a los médicos que no tienen intereses en los pagadores (dos órdenes de magnitud superior).
- En general, los médicos que tienen intereses en los pagadores reciben entorno al 43% de la cantidad invertida, por lo que están lejos de recuperar la cantidad invertida mediante pagos recibidos.

Figura 17 - Indicadores numéricos relativos al análisis de los intereses de los médicos

En la Figura 18, se comprueba que la relación entre importes recibidos sobre cantidades invertidas no es homogénea durante los diferentes años. Por ejemplo, en los años 2015 y 2017, el promedio de los importes recibidos es similar (cerca de los 22.000 $) mientras que el promedio de las cantidades invertidas es sensiblemente diferente (unos 80.000 $ en 2015 y unos 120.000 $ en 2017).

Figura 18 - Gráfico de dispersión de importes recibidos vs cantidades invertidas

6. Conclusiones

El reto de diseñar un sistema completo de explotación de los datos de Open Payments que se ha planteado en este trabajo ha sido ejemplo muy realista de una situación que se puede dar en cualquier organización: se dispone de datos sin explotar y se desea desarrollar un sistema para obtener información y conocimiento útil para la organización.

En el presente trabajo, se han realizado tanto las tareas de diseño como de implementación de todos los subsistemas necesarios. Esto ha supuesto un doble desafío, ya que, además cubrir todos los aspectos analíticos y de negocio, se han tenido que resolver cuestiones estrictamente asociadas con la infraestructura tecnológica. Por tanto, este trabajo es una muestra de cómo, en los proyectos que se llevan a cabo en las organizaciones, los expertos analíticos deben trabajar de la mano de los expertos tecnológicos para que las soluciones desarrolladas sean óptimas.

Al finalizar el trabajo, se ha conseguido cumplir todos los objetivos descritos en el apartado 1.2. Para lograrlo, ha sido clave el establecimiento de un plan de trabajo realista y bien definido. La etapa más importante ha sido la de análisis y de diseño preliminar. Ésta se ha dimensionado de manera adecuada para tener tiempo suficiente para obtener unos requisitos claros y una arquitectura global sólida, de modo que en la etapa de diseño detallado e implementación no ha habido cambios que obligasen a una revisión del alcance.

Sin embargo, en la etapa de diseño detallado e implementación sí que ha habido cambios con respecto a lo planificado. En lugar de seguir la metodología en cascada mostrada en el Anexo 8, ha sido inevitable adoptar una estrategia

iterativa más cercana a las metodologías ágiles. En primer lugar, ha sido necesario unir tareas asociadas a varios subsistemas en bloques de trabajo simultáneo, en concreto los subsistemas de almacenamiento y ETL por un lado y los subsistemas de análisis y visualización por otro. En segundo lugar, en lugar de realizar todas las actividades de diseño y, posteriormente, todas las actividades de implementación, se ha optado por encadenar las etapas de diseño e implementación de los dos bloques de trabajo mencionados. Cada una de estas etapas ha consistido en una iteración diseño → implementación → pruebas → rediseño, con tantas repeticiones como han sido necesarias hasta lograr los objetivos planteados. A pesar de estos cambios, la duración real de la etapa de diseño detallado e implementación ha sido la planificada, lo que probablemente no se hubiese conseguido de no adoptar la nueva estrategia.

Por otro lado, durante la fase de implementación se ha concluido que identificar la consideración del caso de uso planteado como big data es más complicado de lo que parece a priori. El análisis de las características de los datos realizada en el apartado 3.1 concluyó que el caso de uso planteado en este trabajo no correspondía a un sistema big data. En cambio, en el momento de la implementación se observó cómo con los recursos disponibles y la solución definida, el volumen y la estructura de los datos almacenados impedían hacer análisis de manera medianamente fluida, teniendo que realizar cambios cercanos a la estrategia “divide y vencerás” característica de los sistemas big data.

Debido a las limitaciones temporales y presupuestarias del trabajo, hay algunos aspectos que no se han podido abordar. Si a esto le sumamos la experiencia adquirida durante la realización del trabajo, se obtienen varias líneas de trabajo que sería interesante desarrollar en el futuro.

- Se debería trabajar con personas conocedoras de la industria sanitaria para mejorar la información obtenida a partir del sistema de explotación creado en este trabajo. De este modo, se aportaría mayor valor de negocio.
- Se debería mejorar la infraestructura que soporta el sistema de explotación diseñado. Ha quedado patente que la solución implementada en un ordenador personal no es suficiente, por lo que habría que estudiar la migración a infraestructura con mayor capacidad. Probablemente, la migración a un cloud público sería una opción para hacerlo a un coste relativamente bajo.
- Se debería profundizar en la limpieza de ciertos datos. Por ejemplo, en la elaboración del trabajo no se han utilizado los nombres de los productos debido a que necesitaban mucho preprocesado al tener erratas, multitud de formatos y tamaños, etc. También se ha observado que los nombres de las ciudades destino de los viajes pagados por el programa tenían muchas erratas e información incoherente. Mejorar los procesos de limpieza aportaría más información y de mayor calidad.

7. Bibliografía

[1] PwC, «Diez temas candentes de la Sanidad Española para 2013,» 2013. [En línea]. Available: https://www.pwc.es/es/publicaciones/sector-publico/assets/diez-temas-candentes-sanidad-2013.pdf (pág.85).

[2] Farmaindustria, «Sistema de autoregulación - Farmaindustria ¿Qué es?,» [En línea]. Available: https://www.codigofarmaindustria.org/servlet/sarfi/quees.html.

[3] CMS - Centers for Medicare & Medicaid Services, «Open Payments Data CMS - About,» [En línea]. Available: https://openpaymentsdata.cms.gov/about.

[4] Centers for Medicare and Medicaid Services, «Dataset Downloads,» [En línea]. Available: https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html.

[5] CMS - Centers for Medicare & Medicaid Services, «Open Payments Public Use Files: Methodology Overview & Data Dictionary,» [En línea]. Available: https://www.cms.gov/OpenPayments/Downloads/OpenPaymentsDataDictionary.pdf.

[6] Oracle, «MySQL,» [En línea]. Available: https://www.mysql.com/.

[7] Talend, «Talend Open Studio for Data Integration,» [En línea]. Available: https://es.talend.com/products/data-integration/data-integration-open-studio/.

[8] Qlik, «Qlik Sense,» [En línea]. Available: https://www.qlik.com/us/products/qlik-sense.

[9] Intel, «Procesador Intel® Core™ i5-7200U - Especificaciones de producto,» [En línea]. Available: https://ark.intel.com/es/products/95443/Intel-Core-i5-7200U-Processor-3M-Cache-up-to-3-10-GHz-.

[10] MySQL, «Download MySQL Community Server,» [En línea]. Available: https://dev.mysql.com/downloads/mysql/.

[11] MySQL, «Download MySQL Workbench,» [En línea]. Available: https://dev.mysql.com/downloads/workbench/.

[12] Hassmann Software, «MySQL Workbench lost connection to MySQL server (after 600 seconds),» [En línea]. Available:

https://web.archive.org/web/20130322031356/http://www.hassmann-software.de/mysql-workbench-lost-connection-to-mysql-server-after-600-seconds.

[13] Fasthosts, «Back up and restore MySQL databases using MySQL Workbench 6 or 8,» [En línea]. Available: https://help.fasthosts.co.uk/app/answers/detail/a_id/2133/~/back-up-and-restore-mysql-databases-using-mysql-workbench-6-or-8.

[14] Qlik, «Probar Qlik Sense® Desktop gratis,» [En línea]. Available: https://www.qlik.com/us/try-or-buy/download-qlik-sense.

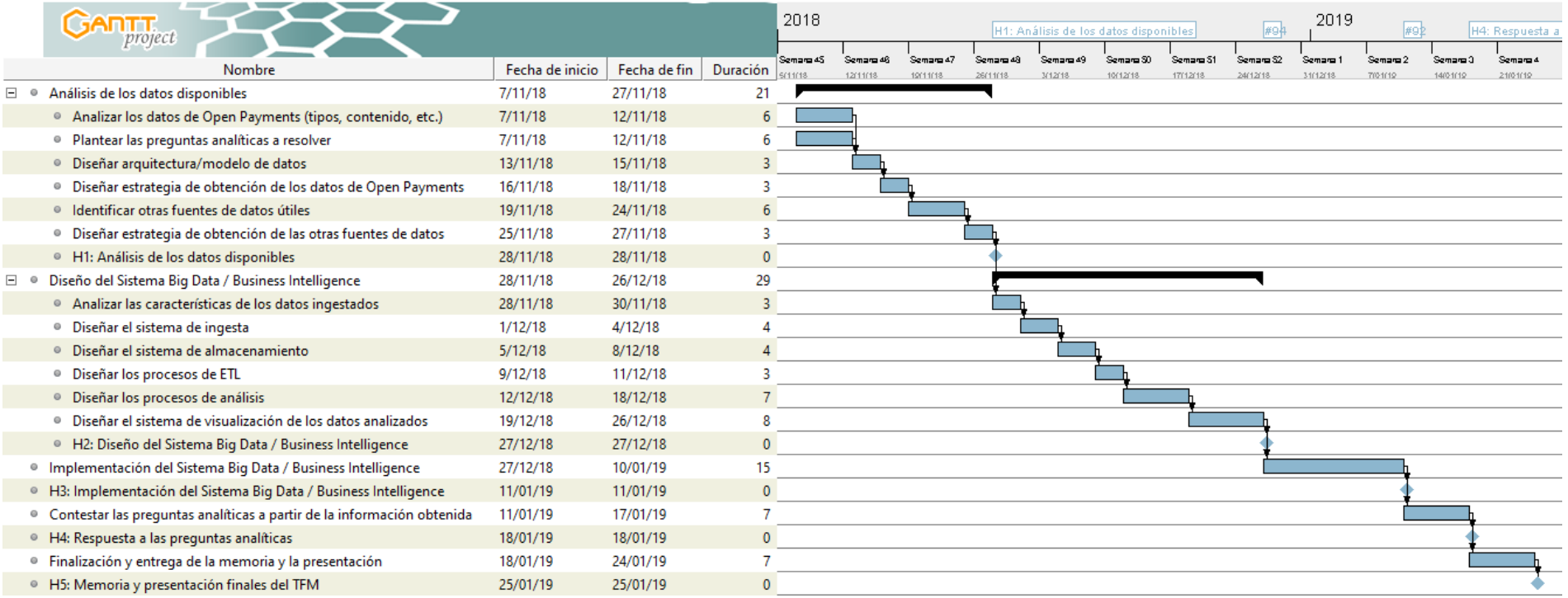
[15] Qlik Community, «Qliksense connector for Mysql Community Edition?,» [En línea]. Available: https://community.qlik.com/t5/New-to-Qlik-Sense/Qliksense-connector-for-Mysql-Community-Edition/td-p/1204267.

[16] MySQL, «Download Connector/ODBC,» [En línea]. Available: https://dev.mysql.com/downloads/connector/odbc/.

[17] Wikipedia, «United States Minor Outlying Islands,» [En línea]. Available: https://en.wikipedia.org/wiki/United_States_Minor_Outlying_Islands.

Anexos

8. Diagrama de Gantt de las tareas



Nombre	Fecha de inicio	Fecha de fin	Duración
Análisis de los datos disponibles	7/11/18	27/11/18	21
Analizar los datos de Open Payments (tipos, contenido, etc.)	7/11/18	12/11/18	6
Plantear las preguntas analíticas a resolver	7/11/18	12/11/18	6
Diseñar arquitectura/modelo de datos	13/11/18	15/11/18	3
Diseñar estrategia de obtención de los datos de Open Payments	16/11/18	18/11/18	3
Identificar otras fuentes de datos útiles	19/11/18	24/11/18	6
Diseñar estrategia de obtención de las otras fuentes de datos	25/11/18	27/11/18	3
H1: Análisis de los datos disponibles	28/11/18	28/11/18	0
Diseño del Sistema Big Data / Business Intelligence	28/11/18	26/12/18	29
Analizar las características de los datos ingestados	28/11/18	30/11/18	3
Diseñar el sistema de ingesta	1/12/18	4/12/18	4
Diseñar el sistema de almacenamiento	5/12/18	8/12/18	4
Diseñar los procesos de ETL	9/12/18	11/12/18	3
Diseñar los procesos de análisis	12/12/18	18/12/18	7
Diseñar el sistema de visualización de los datos analizados	19/12/18	26/12/18	8
H2: Diseño del Sistema Big Data / Business Intelligence	27/12/18	27/12/18	0
Implementación del Sistema Big Data / Business Intelligence	27/12/18	10/01/19	15
H3: Implementación del Sistema Big Data / Business Intelligence	11/01/19	11/01/19	0
Contestar las preguntas analíticas a partir de la información obtenida	11/01/19	17/01/19	7
H4: Respuesta a las preguntas analíticas	18/01/19	18/01/19	0
Finalización y entrega de la memoria y la presentación	18/01/19	24/01/19	7
H5: Memoria y presentación finales del TFM	25/01/19	25/01/19	0

Figura 19 - Planificación de las tareas del TFM

9. Open Payments Methodology Overview & Data Dictionary

En este anexo se presentan algunos extractos de los puntos más relevantes del documento Open Payments Methodology Overview & Data Dictionary [5] en el que se explica la metodología de obtención de datos y el formato de los mismos.

Se mantiene el idioma original y la referencia a las secciones internas del documento. En algún caso se cambia el formato o se añaden anotaciones para facilitar la comprensión.

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 09`.

10. Catalogación de los datos de Open Payments

En la siguiente tabla, se muestran los diferentes ficheros de datos en los que aparece cada dato codificados con el apéndice en el que se explican los diferentes archivos de datos de Open Payments:

- Appendix B: General Payments Detail (Program Year 2016 and Upcoming Years)
- Appendix C: General Payments Detail (Program Years 2013-2015)
- Appendix D: Research Payments Detail (Program Year 2016 and Upcoming Years)
- Appendix E: Research Payments Detail (Program Years 2013-2015)
- Appendix F: Physician Ownership Information Detail (All Program Years)
- Appendix G: Deleted and Removed Records File
- Appendix H: Physician Profile Supplement File

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Change_Type	VARCHAR2(20)	X	X	X	X	X	X		---	---	---
Payment_Type	VARCHAR2(50)						X		---	---	---
Product_Indicator	VARCHAR2(50)	X	X	X					---	---	---
Record_ID	NUMBER(38,0)	X	X	X	X	X	X	X	h_payment	record_id	BIGINT
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	VARCHAR2(38)	X	X	X	X	X		X	h_payment	applicable_payer_id	BIGINT
Date_of_Payment	DATE	X	X	X	X				h_payment	date_of_payment	DATE
Form_of_Payment_or_Transfer_of_Value	VARCHAR2(100)	X	X	X	X				h_payment	form_of_payment	VARCHAR(100)
Nature_of_Payment_or_Transfer_of_Value	VARCHAR2(200)	X	X						h_payment	nature_of_payment	VARCHAR(200)
Number_of_Payments_Included_in_Total_Amount	NUMBER(3,0)	X	X						h_payment	number_of_payments_included	INT
Program_Year	CHAR(4)	X	X	X	X	X			h_payment	program_year	INT
Total_Amount_of_Payment_USDollars	NUMBER(12,2)	X	X	X	X				h_payment	total_amount_of_payment_usdollars	FLOAT
Delay_in_Publication_Indicator	CHAR(3)	X	X	X	X				h_payment	delay_in_publication_indicator	CHAR(3)
Dispute_Status_for_Publication	CHAR(3)	X	X	X	X	X			h_payment	dispute_status_for_publication	CHAR(3)
Payment_Publication_Date	DATE	X	X	X	X	X			h_payment	payment_publication_date	DATE
Covered_Recipient_Type	VARCHAR2(50)	X	X	X	X				h_payment	covered_recipient_type	VARCHAR(50)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Physician_Profile_ID	NUMBER(38,0)	X	X	X	X	X		X	h_payment	physician_profile_id	BIGINT
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	h_payment	investigator_1_profile_id	BIGINT
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	h_payment	investigator_2_profile_id	BIGINT
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	h_payment	investigator_3_profile_id	BIGINT
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	h_payment	investigator_4_profile_id	BIGINT
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	h_payment	investigator_5_profile_id	BIGINT
Teaching_Hospital_ID	NUMBER(38,0)	X	X	X	X			X	h_payment	teaching_hospital_id	BIGINT
Noncovered_Recipient_Entity_Name	VARCHAR2(50)			X	X				h_payment	noncovered_recipient_entity_name	VARCHAR(50)
Recipient_Primary_Business_Street_Address_Line1	VARCHAR2(55)	X	X	X	X	X			h_payment	recipient_primary_business_street_address_line1	VARCHAR(55)
Recipient_Primary_Business_Street_Address_Line2	VARCHAR2(55)	X	X	X	X	X			h_payment	recipient_primary_business_street_address_line2	VARCHAR(55)
Recipient_City	VARCHAR2(40)	X	X	X	X	X			h_payment	recipient_city	VARCHAR(40)
Recipient_Postal_Code	VARCHAR2(20)	X	X	X	X	X			h_payment	recipient_postal_code	VARCHAR(20)
Recipient_Zip_Code	VARCHAR2(10)	X	X	X	X	X			h_payment	recipient_zip_code	VARCHAR(10)
Recipient_Province	VARCHAR2(20)	X	X	X	X	X			h_payment	recipient_province	VARCHAR(20)
Recipient_State	CHAR(2)	X	X	X	X	X			h_payment	recipient_state	CHAR(2)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Recipient_Country	VARCHAR2(100)	X	X	X	X	X			h_payment	recipient_country	VARCHAR(100)
Related_Product_Indicator	VARCHAR2(100)	X		X					h_payment	related_product_indicator	CHAR(3)
City_of_Travel	VARCHAR2(40)	X	X						h_payment	city_of_travel	VARCHAR(40)
Country_of_Travel	VARCHAR2(100)	X	X						h_payment	country_of_travel	VARCHAR(100)
State_of_Travel	CHAR(2)	X	X						h_payment	state_of_travel	CHAR(2)
Preclinical_Research_Indicator	CHAR(3)			X	X				h_payment	preclinical_research_indicator	CHAR(3)
ClinicalTrials_Gov_Identifier	VARCHAR2(11)			X	X				h_payment	clinicaltrials_gov_identifier	VARCHAR(11)
Context_of_Research	VARCHAR2(500)			X	X				h_payment	context_of_research	VARCHAR(500)
Expenditure_Category1	VARCHAR2(50)			X	X				h_payment	expenditure_category1	VARCHAR(50)
Expenditure_Category2	VARCHAR2(50)			X	X				h_payment	expenditure_category2	VARCHAR(50)
Expenditure_Category3	VARCHAR2(50)			X	X				h_payment	expenditure_category3	VARCHAR(50)
Expenditure_Category4	VARCHAR2(50)			X	X				h_payment	expenditure_category4	VARCHAR(50)
Expenditure_Category5	VARCHAR2(50)			X	X				h_payment	expenditure_category5	VARCHAR(50)
Expenditure_Category6	VARCHAR2(50)			X	X				h_payment	expenditure_category6	VARCHAR(50)
Name_of_Study	VARCHAR2(500)			X	X				h_payment	name_of_study	VARCHAR(500)
Research_Information_Link	VARCHAR2(2083)			X	X				h_payment	research_information_link	VARCHAR(2083)
Interest_Held_by_Physician_or_an_Immediate_Family_Member	VARCHAR2(50)					X			h_payment	interest_held_by_physician	VARCHAR(50)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Physician_Ownership_Indicator	CHAR(3)	X	X						h_payment	physician_ownership_indicator	CHAR(3)
Terms_of_Interest	VARCHAR2(500)					X			h_payment	terms_of_interest	VARCHAR(500)
Total_Amount_Invested_USDollars	NUMBER(12,2)					X			h_payment	total_amount_invested_usdollars	FLOAT
Value_of_Interest	NUMBER(12,2)					X			h_payment	value_of_interest	FLOAT
Charity_Indicator	CHAR(3)	X	X						h_payment	charity_indicator	CHAR(3)
Name_of_Third_Party_Entity_Receiving_Payment_or_Transfer_of_Value	VARCHAR2(50)	X	X						h_payment	name_of_third_party_entity_receiving_payment	VARCHAR(50)
Third_Party_Equals_Covered_Recipient_Indicator	CHAR(3)	X	X						h_payment	third_party_equals_covered_recipient_indicator	CHAR(3)
Third_Party_Payment_Recipient_Indicator	VARCHAR2(50)	X	X						h_payment	third_party_payment_recipient_indicator	VARCHAR(50)
Contextual_Information	VARCHAR2(500)	X	X						h_payment	contextual_information	VARCHAR(500)
Record_ID	NUMBER(38,0)	X	X	X	X	X	X	X	h_payment_product	record_id	BIGINT
Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_X (1)	VARCHAR2(100)	X	X	X	X			X	h_payment_product	product_name	VARCHAR(100)
Teaching_Hospital_ID	NUMBER(38,0)	X	X	X	X			X	d_hospital	teaching_hospital_id	BIGINT
Teaching_Hospital_CCN	VARCHAR2(06)	X	X	X	X				d_hospital	teaching_hospital_ccn	VARCHAR(6)
Teaching_Hospital_Name	VARCHAR2(100)	X	X	X	X				d_hospital	teaching_hospital_name	VARCHAR(100)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Principal_Investigator_X_Profile_ID (1)	NUMBER(38,0)			X	X			X	d_investigator	investigator_profile_id	BIGINT
Principal_Investigator_X_First_Name (1)	VARCHAR2(20)			X	X				d_investigator	investigator_first_name	VARCHAR(20)
Principal_Investigator_X_Middle_Name (1)	VARCHAR2(20)			X	X				d_investigator	investigator_middle_name	VARCHAR(20)
Principal_Investigator_X_Last_Name (1)	VARCHAR2(35)			X	X				d_investigator	investigator_last_name	VARCHAR(35)
Principal_Investigator_X_Name_Suffix_ (1)	VARCHAR2(5)			X	X				d_investigator	investigator_name_suffix	VARCHAR(5)
Principal_Investigator_X_Primary_Type (1)	VARCHAR2(50)			X	X				d_investigator	investigator_primary_type	VARCHAR(50)
Principal_Investigator_X_Specialty (1)	VARCHAR2(300)			X	X				d_investigator	investigator_specialty	VARCHAR(300)
Principal_Investigator_X_Business_Street_Address_Line1 (1)	VARCHAR2(55)			X	X				d_investigator	investigator_business_street_address_line1	VARCHAR(55)
Principal_Investigator_X_Business_Street_Address_Line2 (1)	VARCHAR2(55)			X	X				d_investigator	investigator_business_street_address_line2	VARCHAR(55)
Principal_Investigator_X_City (1)	VARCHAR2(40)			X	X				d_investigator	investigator_city	VARCHAR(40)
Principal_Investigator_X_Postal_Code (1)	VARCHAR2(20)			X	X				d_investigator	investigator_postal_code	VARCHAR(20)
Principal_Investigator_X_Zip_Code (1)	VARCHAR2(10)			X	X				d_investigator	investigator_zip_code	VARCHAR(10)
Principal_Investigator_X_Province (1)	VARCHAR2(20)			X	X				d_investigator	investigator_province	VARCHAR(20)
Principal_Investigator_X_State (1)	CHAR(2)			X	X				d_investigator	investigator_state	CHAR(2)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Principal_Investigator_X_Country (1)	VARCHAR2(100)			X	X				d_investigator	investigator_country	VARCHAR(100)
Principal_Investigator_X_License_State_code1 (1)	CHAR(2)			X	X				d_investigator	investigator_license_state_code1	CHAR(2)
Principal_Investigator_X_License_State_code2 (1)	CHAR(2)			X	X				d_investigator	investigator_license_state_code2	CHAR(2)
Principal_Investigator_X_License_State_code3 (1)	CHAR(2)			X	X				d_investigator	investigator_license_state_code3	CHAR(2)
Principal_Investigator_X_License_State_code4 (1)	CHAR(2)			X	X				d_investigator	investigator_license_state_code4	CHAR(2)
Principal_Investigator_X_License_State_code5 (1)	CHAR(2)			X	X				d_investigator	investigator_license_state_code5	CHAR(2)
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	VARCHAR2(38)	X	X	X	X	X		X	d_payer	applicable_payer_id	BIGINT
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	VARCHAR2(100)	X	X	X	X	X			d_payer	applicable_payer_name	VARCHAR(100)
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	CHAR(2)	X	X	X	X	X			d_payer	applicable_payer_state	CHAR(2)
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	VARCHAR2(100)	X	X	X	X	X			d_payer	applicable_payer_country	VARCHAR(100)
Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	VARCHAR2(100)	X	X	X	X	X			d_payer	submitting_payer_name	VARCHAR(100)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Physician_Profile_ID	NUMBER(38,0)	X	X	X	X	X		X	d_physician	physician_profile_id	BIGINT
Physician_First_Name	VARCHAR2(20)	X	X	X	X	X			d_physician	physician_first_name	VARCHAR(20)
Physician_Middle_Name	VARCHAR2(20)	X	X	X	X	X			d_physician	physician_middle_name	VARCHAR(20)
Physician_Last_Name	VARCHAR2(35)	X	X	X	X	X			d_physician	physician_last_name	VARCHAR(35)
Physician_Name_Suffix	VARCHAR2(5)	X	X	X	X	X			d_physician	physician_name_suffix	VARCHAR(5)
Physician_Primary_Type	VARCHAR2(100)	X	X	X	X	X			d_physician	physician_primary_type	VARCHAR(100)
Physician_Specialty	VARCHAR2(300)	X	X	X	X	X			d_physician	physician_specialty	VARCHAR(300)
Physician_License_State_code1	CHAR(2)	X	X	X	X				d_physician	physician_license_state_code1	CHAR(2)
Physician_License_State_code2	CHAR(2)	X	X	X	X				d_physician	physician_license_state_code2	CHAR(2)
Physician_License_State_code3	CHAR(2)	X	X	X	X				d_physician	physician_license_state_code3	CHAR(2)
Physician_License_State_code4	CHAR(2)	X	X	X	X				d_physician	physician_license_state_code4	CHAR(2)
Physician_License_State_code5	CHAR(2)	X	X	X	X				d_physician	physician_license_state_code5	CHAR(2)
Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_X (1) (2)	VARCHAR2(100)	X	X	X	X			X	d_product	product_name	VARCHAR(100)
Covered_or_Noncovered_Indicator_X (1)	VARCHAR2(100)	X		X					d_product	covered_or_noncovered_indicator	VARCHAR(100)
Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_X (1)	VARCHAR2(100)	X		X					d_product	product_type	VARCHAR(100)

Original field name	Original data type	B	C	D	E	F	G	DB	Table	Field	Data type
Product_Category_or_Therapeutic_Area_X (1)	VARCHAR2(100)	X		X					d_product	product_category_or_therapeutic_area	VARCHAR(100)
Associated_Drug_or_Biological_NDC_X (1) (2)	VARCHAR2(100)	X	X	X	X				d_product	associated_drug_or_biological_ndc	VARCHAR(100)

Tabla 10 - Catalogación de los datos de Open Payments

11. Script de implementación del modelo de datos

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 11`.

12. Formatos de las conexiones con la base de datos y de los CSV en Talend Open Studio

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 12`.

13. Trabajos de los procesos de ETL

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 13`.

14. Script de fragmentación de ficheros CSV

```
import os

def split(filehandler, row_limit=10000,
          output_name_template='output_%s.csv', output_path='.', keep_headers=True):

    current_piece = 1
    current_limit = row_limit
    current_out_path = os.path.join(
        output_path,
        output_name_template % current_piece
    )
    current_out_writer = open(current_out_path, 'wb')

    i = 0
    with filehandler as f:
        for line in f:
            if i == 0:
                header = line
                current_out_writer.write(header.encode('utf8'))
            elif i + 1 <= current_limit:
                current_out_writer.write(line.encode('utf8'))
            else:
                current_piece += 1
                current_limit = row_limit * current_piece
                current_out_path = os.path.join(
                    output_path,
                    output_name_template % current_piece
                )
                current_out_writer.close()
                current_out_writer = open(current_out_path, 'wb')
                if keep_headers:
                    current_out_writer.write(header.encode('utf8'))
                current_out_writer.write(line.encode('utf8'))
            i += 1

ruta = 'C:\\Users\\ratan\\Desktop\\RAUL\\datos'
f_in = 'OP_DTL_GNRL_PGYR2017_P06292018'
ext_in = '.csv'
f_out = f_in + '_%02d' + ext_in
split(open(ruta + '\\' + f_in + ext_in, 'r', encoding='utf-8'),
      row_limit=2100000, output_name_template=f_out, output_path=ruta)
```

15. Tiempos de carga inicial de datos

La siguiente tabla muestra los tiempos de carga iniciales de los datos de Open Payments.

Año	Trabajo	Inicio	Fin	Duración (hh:mm)	Num. de pagos	Pagos acumulados	Num. de errores
2013	1.1	19/12/2018 19:20	20/12/2018 00:47	05:27	2,099,999	2,099,999	0
2013	1.2	20/12/2018 00:56	20/12/2018 04:33	03:37	2,071,205	4,171,204	0
2013	2	20/12/2018 09:47	20/12/2018 10:25	00:38	434,683	4,605,887	0
2013	3	20/12/2018 10:26	20/12/2018 10:26	00:00	5,175	4,611,062	0
2013	4	20/12/2018 10:27	20/12/2018 10:27	00:00	0	4,611,062	0
2014	5.1	20/12/2018 10:31	20/12/2018 16:25	05:54	2,099,999	6,711,061	0
2014	5.2	20/12/2018 16:51	20/12/2018 22:45	05:54	2,100,000	8,811,061	0
2014	5.3	20/12/2018 23:51	21/12/2018 05:40	05:49	2,100,000	10,911,061	0
2014	5.4	21/12/2018 09:43	21/12/2018 16:20	06:37	2,100,000	13,011,061	0
2014	5.5	21/12/2018 16:45	21/12/2018 22:25	05:40	2,100,000	15,111,061	0
2014	5.6	21/12/2018 22:28	22/12/2018 00:19	01:51	806,929	15,917,990	0
2014	6	22/12/2018 00:23	22/12/2018 02:03	01:40	730,138	16,648,128	0
2014	7	22/12/2018 02:04	22/12/2018 02:05	00:01	5,370	16,653,498	0
2014	8	22/12/2018 02:06	22/12/2018 02:06	00:00	0	16,653,498	0
2015	9.1	22/12/2018 02:08	22/12/2018 05:57	03:49	2,099,999	18,753,497	0
2015	9.2	22/12/2018 09:49	22/12/2018 13:56	04:07	2,100,000	20,853,497	0
2015	9.3	22/12/2018 14:13	22/12/2018 19:07	04:54	2,099,995	22,953,492	5
2015	9.4	22/12/2018 19:18	22/12/2018 22:48	03:30	2,100,000	25,053,492	0
2015	9.6	22/12/2018 23:05	23/12/2018 01:38	02:33	1,044,636	26,098,128	0
2015	9.5	23/12/2018 01:41	23/12/2018 06:05	04:24	2,099,997	28,198,125	3
2015	10	23/12/2018 11:09	23/12/2018 12:45	01:36	870,218	29,068,343	57
2015	11	23/12/2018 12:57	23/12/2018 12:57	00:00	4,718	29,073,061	0
2015	12	23/12/2018 12:59	23/12/2018 12:59	00:00	0	29,073,061	0
2016	13.1	23/12/2018 13:01	23/12/2018 17:16	04:15	2,099,992	31,173,053	7
2016	13.2	23/12/2018 17:56	23/12/2018 22:40	04:44	2,099,999	33,273,052	1

Año	Trabajo	Inicio	Fin	Duración (hh:mm)	Num. de pagos	Pagos acumulados	Num. de errores
2016	13.3	23/12/2018 22:43	24/12/2018 03:23	04:40	2,099,991	35,373,043	9
2016	13.4	24/12/2018 07:04	24/12/2018 12:20	05:16	2,099,994	37,473,037	6
2016	13.6	24/12/2018 14:57	24/12/2018 17:21	02:24	1,159,502	38,632,539	2
2016	14	24/12/2018 18:13	24/12/2018 19:14	01:01	735,184	39,367,723	4
2016	13.5	24/12/2018 19:18	24/12/2018 23:29	04:11	2,100,000	41,467,723	0
2016	15	25/12/2018 01:34	25/12/2018 01:34	00:00	3,785	41,471,508	0
2016	16	25/12/2018 01:36	25/12/2018 01:36	00:00	0	41,471,508	0
2017	17.1	25/12/2018 01:39	25/12/2018 05:27	03:48	2,099,999	43,571,507	0
2017	17.2	25/12/2018 10:26	25/12/2018 14:26	04:00	2,099,991	45,671,498	5
2017	17.3	25/12/2018 15:23	25/12/2018 20:11	04:48	2,100,000	47,771,498	0
2017	17.6	25/12/2018 20:28	25/12/2018 21:25	00:57	431,832	48,203,330	1
2017	18	25/12/2018 21:45	25/12/2018 22:30	00:45	607,839	48,811,169	26
2017	17.4	25/12/2018 23:14	26/12/2018 03:50	04:36	2,099,998	50,911,167	2
2017	17.5	26/12/2018 10:12	26/12/2018 15:17	05:05	2,100,000	53,011,167	0
2017	19	26/12/2018 17:25	26/12/2018 17:25	00:00	2,630	53,013,797	0

Tabla 11 - Tiempos de carga inicial de los datos de Open Payments

En resumen, el proceso de carga:

- Ha tardado 4 días, 13 horas y 27 minutos en procesar y guardar los datos en la base de datos (7 días de trabajo, aproximadamente).
- Ha procesado 53,013,925 pagos.
- Ha encontrado 128 incoherencias de formato en registros de pago, con lo que estos pagos se han desechado.

16. Script de carga de las tablas simplificadas de h_payment

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 16`.

17. Script de cálculo de indicadores

Debido a la extensión de este anexo, se adjunta en el documento adicional `raul caro - memoria tfm - anexo 17`.

18. Tableros de Qlik Sense

18.1. Análisis por tipo de pago

Figura 20 - Análisis por tipo de pago

18.2. Análisis por localización

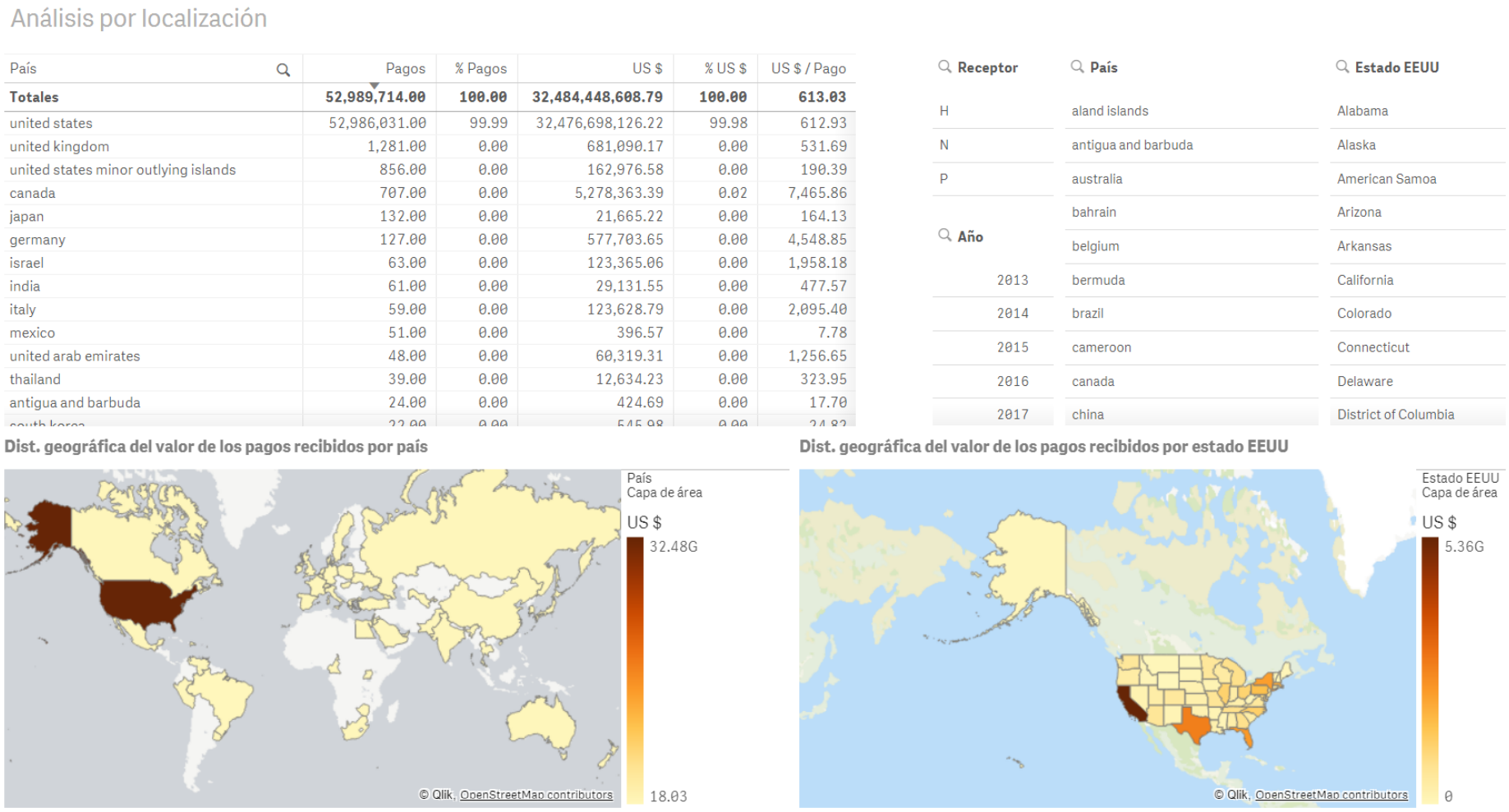
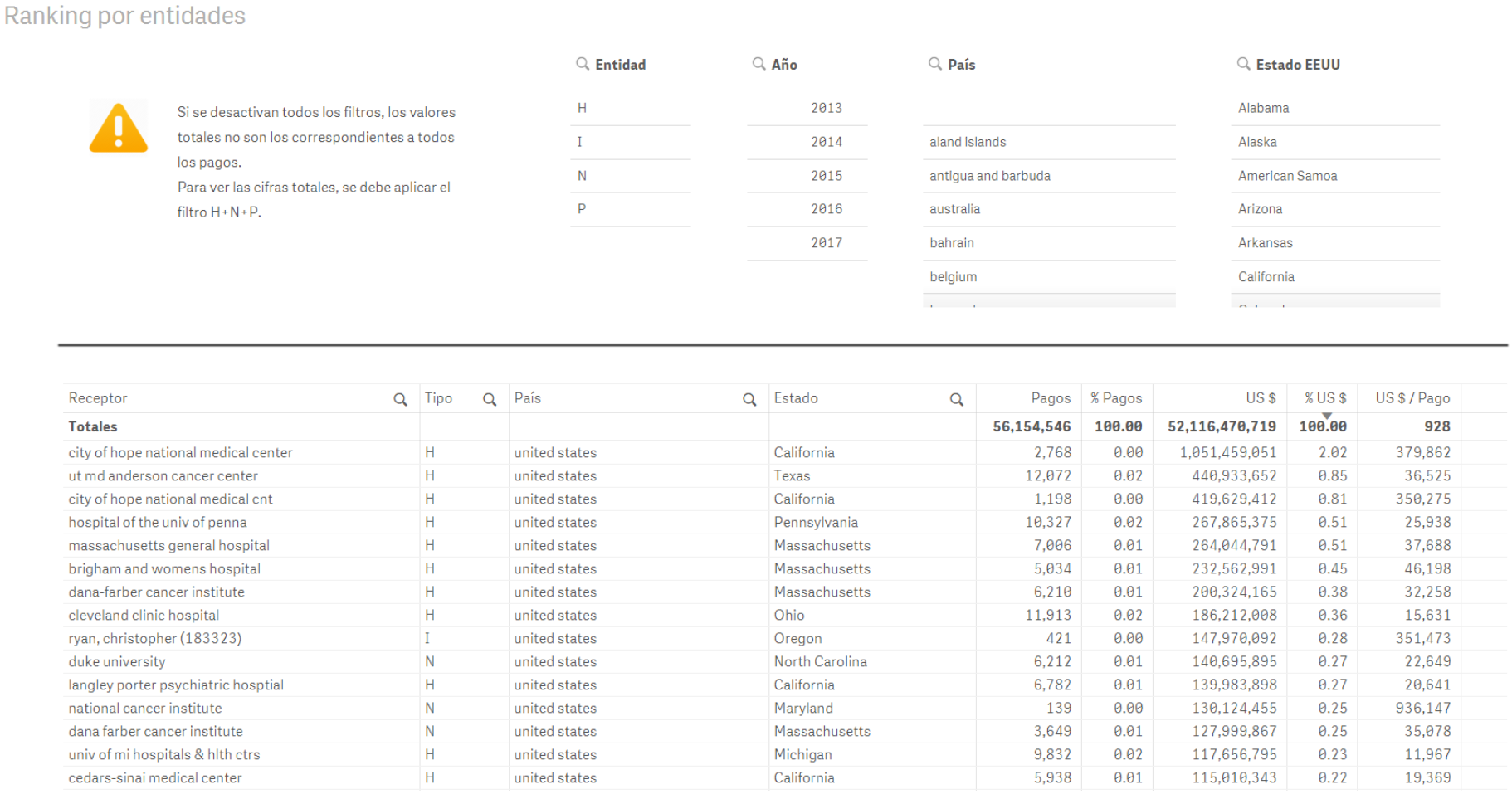


Figura 21 - Análisis por localización

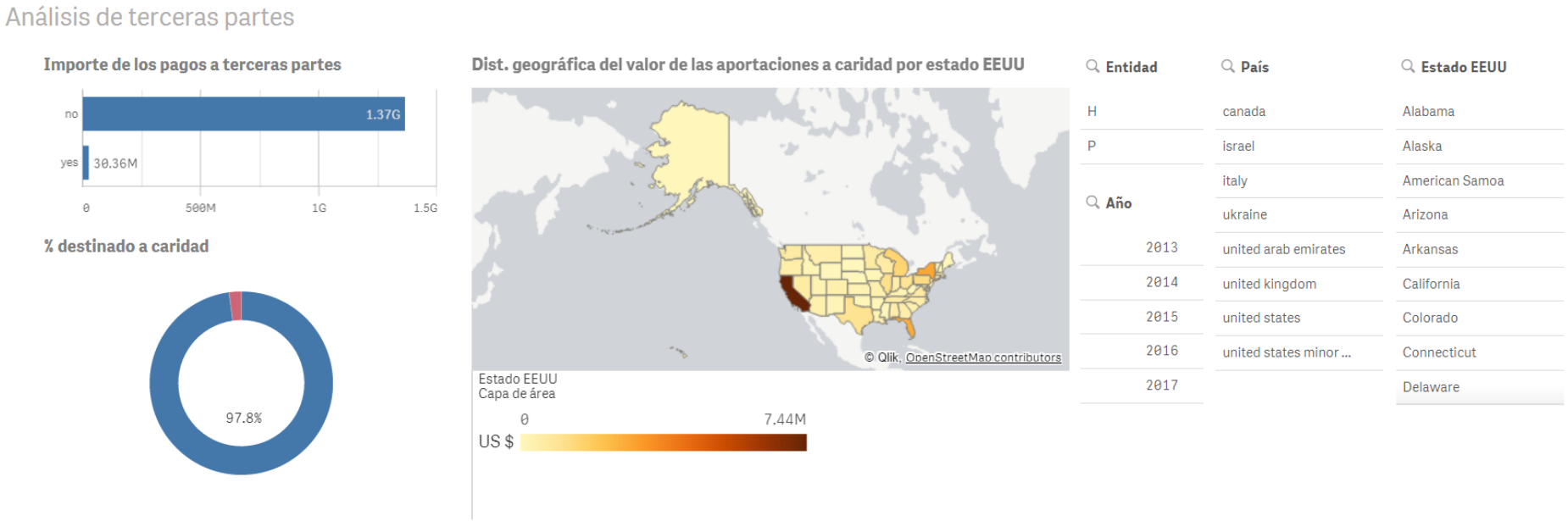
18.3. Ranking por entidades



Receptor	Tipo	País	Estado	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales				**56,154,546**	**100.00**	**52,116,470,719**	**100.00**	**928**
city of hope national medical center	H	united states	California	2,768	0.00	1,051,459,051	2.02	379,862
ut md anderson cancer center	H	united states	Texas	12,072	0.02	440,933,652	0.85	36,525
city of hope national medical cnt	H	united states	California	1,198	0.00	419,629,412	0.81	350,275
hospital of the univ of penna	H	united states	Pennsylvania	10,327	0.02	267,865,375	0.51	25,938
massachusetts general hospital	H	united states	Massachusetts	7,006	0.01	264,044,791	0.51	37,688
brigham and womens hospital	H	united states	Massachusetts	5,034	0.01	232,562,991	0.45	46,198
dana-farber cancer institute	H	united states	Massachusetts	6,210	0.01	200,324,165	0.38	32,258
cleveland clinic hospital	H	united states	Ohio	11,913	0.02	186,212,008	0.36	15,631
ryan, christopher (183323)	I	united states	Oregon	421	0.00	147,970,092	0.28	351,473
duke university	N	united states	North Carolina	6,212	0.01	140,695,895	0.27	22,649
langley porter psychiatric hosptial	H	united states	California	6,782	0.01	139,983,898	0.27	20,641
national cancer institute	N	united states	Maryland	139	0.00	130,124,455	0.25	936,147
dana farber cancer institute	N	united states	Massachusetts	3,649	0.01	127,999,867	0.25	35,078
univ of mi hospitals & hlth ctrs	H	united states	Michigan	9,832	0.02	117,656,795	0.23	11,967
cedars-sinai medical center	H	united states	California	5,938	0.01	115,010,343	0.22	19,369

Figura 22 - Ranking por entidades

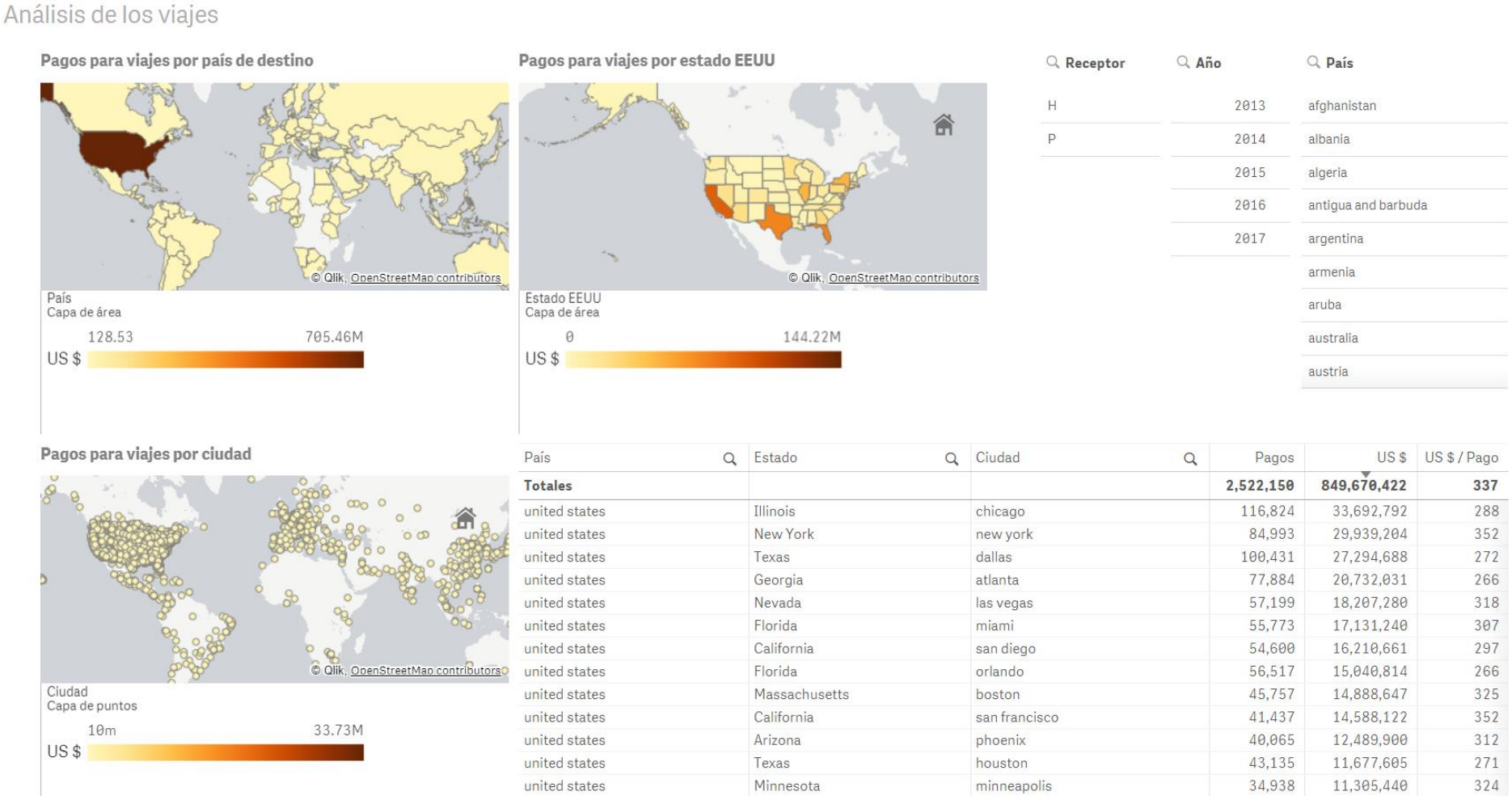
18.4. Análisis de terceras partes



Tercera parte	Receptor	País	Estado	Caridad	Pagos	% Pagos	US $	% US $	US $ / Pago
Totales					**619,116**	**100.00**	**1,399,274,498**	**100.00**	**2,260**
burkhart resource, ltd	P	united states	Texas	no	62	0.01	87,761,169	6.27	1,415,503
guidance endodontics llc	P	united states	New Mexico	no	1	0.00	22,880,194	1.64	22,880,194
mayo foundation	P	united states	Minnesota	no	256	0.04	22,646,567	1.62	88,463
hackensack university medical center foundation	H	united states	New Jersey	no	2	0.00	15,004,000	1.07	7,502,000
james g berbee and karen a walsh joint revocable t	P	united states	Wisconsin	no	3	0.00	14,265,661	1.02	4,755,220
concept properties llc	P	united states	Kentucky	no	56	0.01	13,894,611	0.99	248,118
sms trust	P	united states	Missouri	no	36	0.01	12,411,856	0.89	344,774
lgl spine llc	P	united states	New York	no	12	0.00	12,397,122	0.89	1,033,094
engh consulting inc	P	united states	Virginia	no	18	0.00	11,946,224	0.85	663,679

Figura 23 - Análisis de terceras partes

18.5. Análisis de los viajes



País	Estado	Ciudad	Pagos	US $	US $ / Pago
Totales			**2,522,150**	**849,670,422**	**337**
united states	Illinois	chicago	116,824	33,692,792	288
united states	New York	new york	84,993	29,939,204	352
united states	Texas	dallas	100,431	27,294,688	272
united states	Georgia	atlanta	77,884	20,732,031	266
united states	Nevada	las vegas	57,199	18,207,280	318
united states	Florida	miami	55,773	17,131,240	307
united states	California	san diego	54,600	16,210,661	297
united states	Florida	orlando	56,517	15,040,814	266
united states	Massachusetts	boston	45,757	14,888,647	325
united states	California	san francisco	41,437	14,588,122	352
united states	Arizona	phoenix	40,065	12,489,900	312
united states	Texas	houston	43,135	11,677,605	271
united states	Minnesota	minneapolis	34,938	11,305,440	324

Figura 24 - Análisis de los viajes

18.6. Análisis de los intereses de los médicos

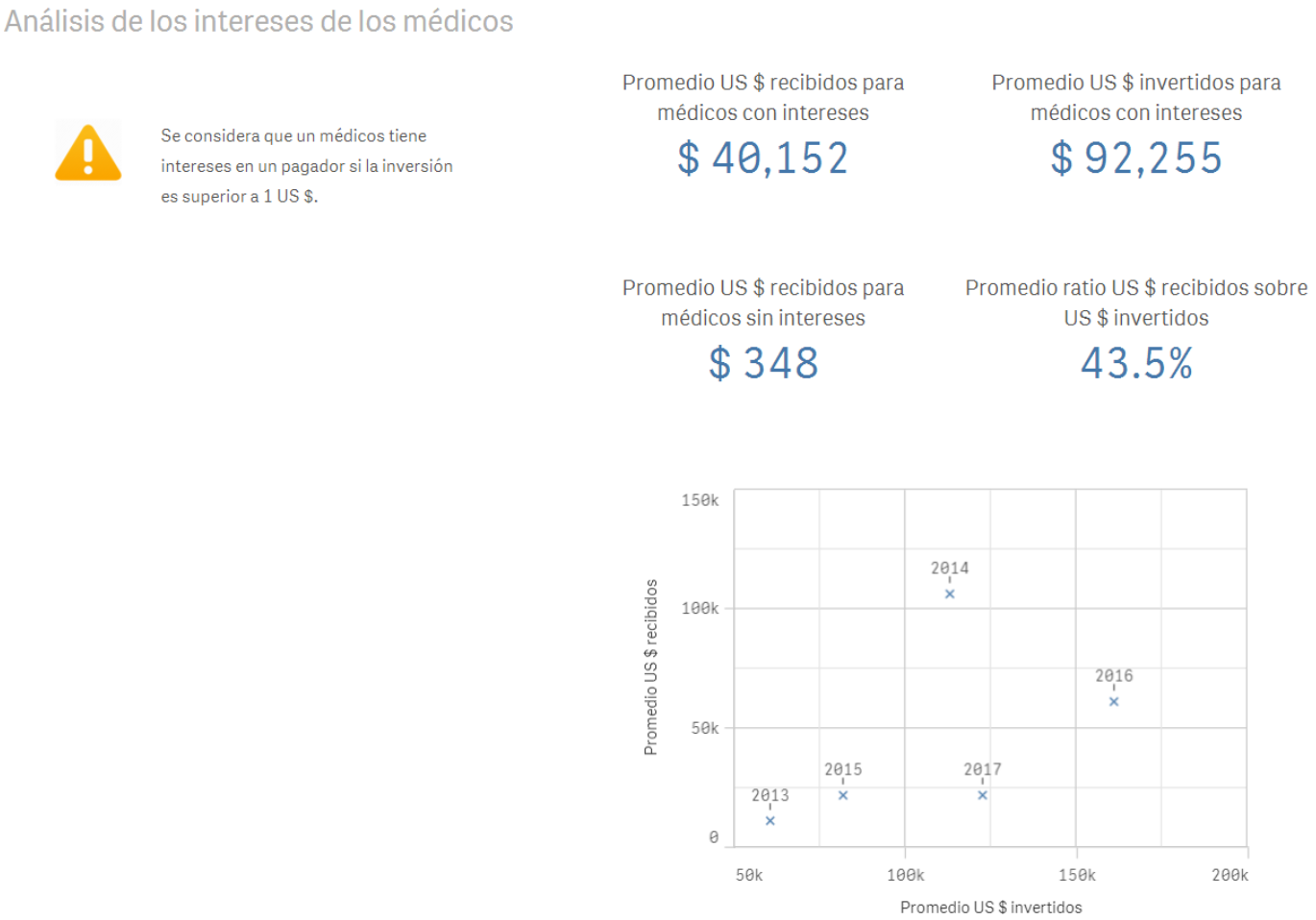


Figura 25 - Análisis de los intereses de los médicos