Construcción de un data warehouse sobre datos de cáncer en Estados Unidos

César Calatrava Ruiz
Ciudad Real, España
Email: cesar.calatrava.ruiz@gmail.com

***Abstract*—**Una de las enfermedades más mortales de hoy en da es el cáncer, cobrándose millones de vctimas, por lo que existe una gran preocupación en el campo médico sobre este tema.
Por todo esto el objetivo de este trabajo consiste en la construcción de un data warehouse que sirva de almacén a un conjunto de datos referentes al cáncer en EE.UU. Dicha construcción se dividirá principalmente en el diseño del modelo y la parte relacionada con el proceso ETL para la extracción, transformación y carga de los datos en el data warehouse.

Keywords*—cáncer, Estados Unidos, data warehouse, proceso ETL, base de datos, buisness intelligence.*

I. Introducción

La salud, actualmente, es un tema de gran calado en la sociedad, y el cáncer uno de sus más claros exponentes, debido al escaso porcentaje de pacientes que consiguen superarlo y recuperarse con éxito.

Algunos datos extrados de la Organización Mundial de la Salud (OMS) corroboran la gran dureza de esta enfermedad. Es una de las primeras causas de muerte a nivel mundial; en 2012 se le atribuyeron 8,2 millones de muertes [1]. No todo los tipos de cáncer tienen la misma importancia en cuanto al número de muertes, los más relevantes son los de pulmón, hgado, estómago, colon y mama. De la misma manera, no afectan igual a hombres y mujeres.

La gran diversidad sociocultural y territorial de los Estados Unidos de América nos lleva a centrar nuestro estudio en esta zona del mundo.

El objetivo de este trabajo es la construcción de un almacén de datos, data warehouse, con la principal meta de almacenar una cantidad considerable de información, en este trabajo serán datos referentes al cáncer, para poder analizarla y dar respuesta a una serie de necesidades no cubiertas con anterioridad.

Para llevar a cabo dicho trabajo, se cuenta con un gran conjunto de datos que recoge las incidencias y muertes por cáncer en EE.UU. a nivel estatal durante diez años, concretamente de 1999 a 2009. La construcción del data warehouse girará alrededor de este *dataset*.

Desde el punto de vista temporal podemos decir que el proceso de toma de decisiones ha sufrido una gran evolución en los últimos treinta años. Esta evolución también ha afectado a la información como tal, ya que en la década de los 70-80 nos referamos a datos, a medida que avanzó pasamos a centrarnos en la información, hasta llegar a la década de los 90 con las decisiones. La década de los 70-80 se caracterizó por los formularios de carácter estático y predefinido, todos estos informes que presentaban un control centralizado estaban enfocados a IT. A medida que fue avanzando la década de los 80 los Sistemas de Ayuda a la Decisión (DSS) y las herramientas OLAP fueron tomando un mayor protagonismo, permitiendo la realización de análisis complejos y test de hipótesis, el acceso de los datos a los niveles de mando medio/alto y enfocado a lneas de negocio. Los términos data warehouse y data mining aparecen en la década de los 90 enfocados principalmente a la generación de hipótesis y a las lneas de negocio e IT conjuntamente, no separadas como pasaba anteriormente.

Actualmente la construcción de los almacenes de datos (bases de datos analticas) viene motivada por las necesidades que no cubren las bases de datos operacionales, destinadas la mayora de las ocasiones a operaciones de carácter diario en el mundo empresarial. Por tanto, las ventajas que aporta el uso de un data warehouse son numerosas: análisis inmediato de los resultados, ahorro en costes de producción, gestión de programas de Marketing, mejor control financiero, cohesión de departamentos, reacción de forma rápida ante los posibles cambios que puedan producirse y una mejor relación con el cliente.

Para poder abordar este problema, en primer lugar, se deberá realizar un estudio de los datos para determinar cuales de ellos se van a utilizar en la confección del almacén, una vez hecho esto el siguiente paso será el diseño y la elección del modelo a seguir, para finalmente pasar al desarrollo del proceso ETL, en el cual extraeremos los datos, los transformaremos y por último, los cargaremos en el data warehouse.

Normalmente el proceso de ETL es el que mayor esfuerzo requiere en el conjunto de todos los pasos para la construcción de un almacén de datos, esto viene derivado principalmente de la utilización de diversas fuentes de datos y de la cuestionable calidad de los mismos. Por tanto, los esfuerzos no deben centrarse en la mejora del proceso ETL, sino en la mejora de la calidad de los datos en los sistemas operacionales.

El resto del trabajo se organiza de la siguiente forma: la sección II describe los datos utilizados de forma breve, por otra parte, la sección III recoge el proceso de diseño del modelo para el almacén, en la sección IV se describirá todo el proceso ETL y la sección V recogerá las diversas aplicaciones del data warehouse. Por último, en la sección VI se expondrán las conclusiones.

II. Descripción de los datos utilizados

A continuación, se dará una explicación más detallada de los datos disponibles con los que se va a trabajar.

Como ya se comentó con anterioridad el conjunto de datos del que se dispone recoge el número de incidencias y muertes por cáncer en EE.UU durante el perodo comprendido entre los años 1999-2009. Dicho conjunto cuenta con 671.640 registros formados por 14 campos, siendo los más relevantes los siguientes: estado, número de casos, tipo de caso (incidencia o muerte), población, tipo de cáncer, año, sexo y raza. Además de otros datos referentes a una serie de valores mnimos, máximos y coeficientes referentes a la edad.

Esta BD se divide por estados, mostrando en cada uno de ellos el número de personas que han padecido cáncer y cuantas de ellas no fueron capaces de superar dicha enfermedad.

A simple vista ya se puede intuir cuales serán las posibles medidas y dimensiones candidatas para el diseño y construcción del almacén.

III. DISEÑO

Para realizar el diseño de nuestro modelo debemos conocer antes las distintas fases de las que consta dicho proceso de diseño:

1) Selección de los proceso a modelar.
2) Decidir el nivel de granularidad.
3) Selección de las dimensiones.
4) Determinar los hechos a considerar.

En cuanto a la estructura que sigue el modelo podemos diferenciar entre un modelo basado en estrella o en copo de nieve, para nuestro trabajo el modelo elegido ha sido un modelo en estrella que nos permite la creación de una base de datos con tiempos de respuesta rápidos, la fácil modificación del diseño ante los posibles cambios. Esta base de datos facilita la interacción con las herramientas de acceso.

La arquitectura elegida para el diseño del modelo ha sido MOLAP, implementación OLAP sobre un motor relacional. Esta arquitectura se caracteriza por su escalabilidad, permite el análisis de una gran cantidad de datos, además de no ser necesaria ninguna herramienta OLAP para su tratamiento al estar basada en SQL podemos acceder a ella con cualquier herramienta de *reporting*.

Una vez que hemos definido la estructura y arquitectura, el siguiente paso es definir las dimensiones ya que estas nos darán las puntas de la estrella. Las dimensiones tomadas han sido:

- Dimensión área: contendrá cada uno de los estados registrados en la base de datos; Alabama, Alaska, etc.
- Dimensión raza: como se comento anteriormente la BD contena un campo raza que ha sido tomado como una dimensión.
- Dimensión zona: esta dimensión corresponde a los diversos tipos de cáncer que incluye nuestro conjunto de datos.
- Dimensión año: es una de las dimensiones más importantes, ya que la mayor parte de los data warehouse se caracterizan por su carácter histórico, es decir, aglutinan datos referentes a varios años.
- Dimensión sexo: esta dimensión recogerá los diversos tipos de sexo; hombre, mujer, y hombre y mujer.

Las medidas son los valores numéricos que el usuario desea analizar a partir de las dimensiones establecidas anteriormente.

Ahora, pasaremos a la elección de la medidas que para nuestro modelo serán tres: el número de incidencias, el número de muertes y el número de habitantes totales. Las medidas son siempre numéricas, son almacenadas en la tabla de hechos y las dimensiones que son textuales se almacenan en las tablas de dimensiones.

Por tanto, el diseño del modelo final queda conformado de la siguiente forma, figura 1. Cada tabla de hechos contiene las claves externas, que se relacionan con sus respectivas tablas de dimensiones, y las columnas con los medidas naturales que serán analizadas.

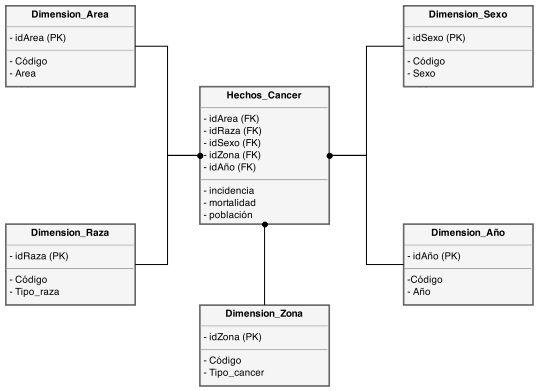

Figura 1. MODELO ROLAP EN ESTRELLA

El motor de bases de datos elegidos para la creación del data warehouse ha sido *SQLite* y como tal donde alojaremos dicho diseño figura 1.

IV. PROCESO ETL

El proceso ETL, que nos permite importar datos desde múltiples fuentes, transformarlos y cargarlos en la base de datos, consta de tres fases bien diferenciadas:

- Extracción: lectura de los datos de las diferentes fuentes de datos.
- Transformación: limpieza y transformación de los datos añadiéndoles contexto y significado.
- Carga: inserción de las tablas de dimensiones y hechos.

a) ***Extracción****:* para esta primera parte del proceso ETL, se ha realizado las extracción de cada una de las dimensiones seleccionadas en la anterior fase: área, raza, zona del cuerpo donde se encuentra el tumor, año y por último el sexo. En esta fase no se ha hecho ningún tratamiento de los datos, dejándolo para el siguiente paso. También se ha hecho la extracción del número de incidencias, muertes y la población total.

Centrándonos más en la parte de desarrollo de esta primera fase, la extracción tanto de las dimensiones como de las medidas ha sido realizada a través de una serie de scripts en Python encargados de seleccionar y almacenar en archivos csv respectivamente los datos.

b) ***Transformación****:* en esta fase los esfuerzos se han centrado en la limpieza y codificación de los datos. Se ha realizado un tratamiento de valores nulos y/o perdidos para las incidencias y muertes en los registros donde en lugar de aparecer la cifra como tal, apareca '?', '-' y/o '.'. Además, se han eliminado los registros referentes al perodo de '2005-2009' ya que no se ha contemplado como parte de la dimensión año. La parte de la limpieza ha sido desarrollada totalmente con varios scripts en Python.

En cuanto a la codificación de las dimensiones, el proceso ha sido el de codificar cada uno de los registros que conforman las tablas de dimensiones, es decir, en la dimensión año por ejemplo al año 1999 se le ha asignado un 0 y as sucesivamente.

Una vez que ya tenemos los datos totalmente limpios y transformados el siguiente paso es la inclusión de los datos referentes a la mortalidad a los demás datos donde ya tenemos las incidencias, población, sexo, etc. Por último, el uso de las claves foráneas (PK) en la tabla de hechos nos lleva a codificar cada uno de los hechos, tal y como se muestra en la figura 2.

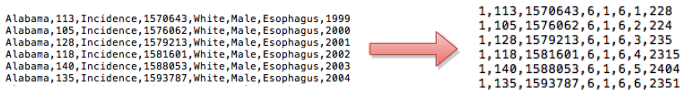

Figura 2. PROCESO DE TRANSFORMACIÓN

Esta última codificación ha sido realizada con las herramientas *rStudio* y *Numbers*.

c) ***Carga****:* en la última fase del proceso ETL se ha realizado la carga de todas la dimensiones y la tabla de hechos que conforman nuestro almacén. Este proceso ha sido realizado con una serie de scripts encargados de ejecutar la sentencia INSER TO en la base de datos de todos los csv donde se recogan las distintas tablas (tabla de hechos y dimensiones).

V. APLICACIONES

A continuación, se expondrán las diferentes aplicaciones que podra tener el data warehouse construido:

1) **Respuesta a una serie de preguntas como:**
   - Cuál es el estado con mayor número de muertes por cáncer de mama en el año 1999?
   - Cuál es el número total de incidencias en el año 2009 por tipo de sexo en Alabama?
   - Cuál es el número total de incidencias ordenadas de forma ascendente en todos los años registrados en función del tipo de cáncer?

   Como podemos observar se trata de una serie de preguntas complejas que no podramos responder de manera trivial sin nuestro data warehouse, en gran parte las de carácter temporal porque como ya se explicó en la introducción una de las caractersticas más reseñable de los almacenes de datos es que almacenan información histórica.

2) **Construcción clasificador bayesiano:** la posibilidad que tenemos de poder calcular la probabilidad que tendrá una persona de sufrir o morir por un determinado cáncer nos lleva a poder realizar la construcción de un clasificador bayesiano (clasificador probabilstico basado en el teorema de Bayes[1]) para intentar estimar la probabilidad que tendrá una persona de padecer cáncer y morir. Por ejemplo, se podra construir un clasificador bayesiano del último año registrado en el data warehouse en el estado de Texas para el cáncer de mama y as se podra estimar la incidencia que tendra en la población dicho cáncer en los siguientes años o poder saber que posibilidad tendra una persona que llega a vivir a Texas.

3) **Generación de mapas de colores:** los mapas de colores dentro del ámbito de la visualización de datos son de gran utilidad y nos permiten de un vistazo rápido tomar perspectiva de los aspectos más destacables en un determinado tema. Con este data warehouse podramos construir estos mapas de colores para conocer de una forma más "visual" la incidencia del cáncer en EE.UU.

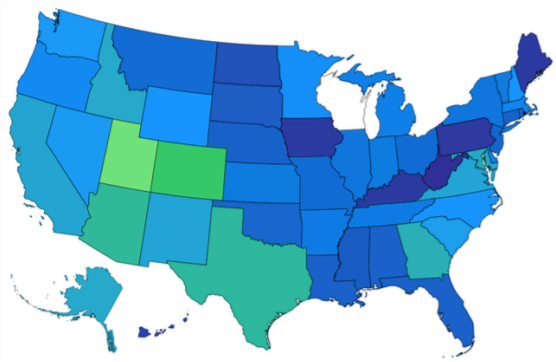

Figura 3. INCIDENCIAS EN COLON AÑO 2009

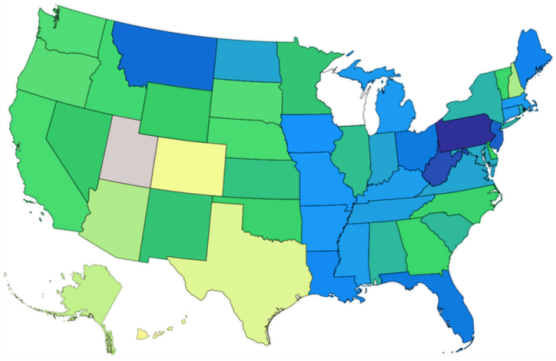

Figura 4. MUERTES EN MAMA AÑO 2009

[1]http://en.wikipedia.org/wiki/Bayesian_probability

En las figuras 3 y 4, se observan dos mapas de colores: el primero recoge número de incidencias por estado en el año 2009 para el cáncer de colon y el segundo referente a la mortalidad estatal por cáncer de mama en el año 2009 también.

4) **Generación de gráficos:** como última aplicación cabe destacar la generación de distintos tipos de gráficos con la herramienta *rStudio*. Con esta herramienta una vez tengamos los datos que nos han sido devueltos al realizar una consulta podremos representarlos gráficamente como en las figuras 5 y 6.

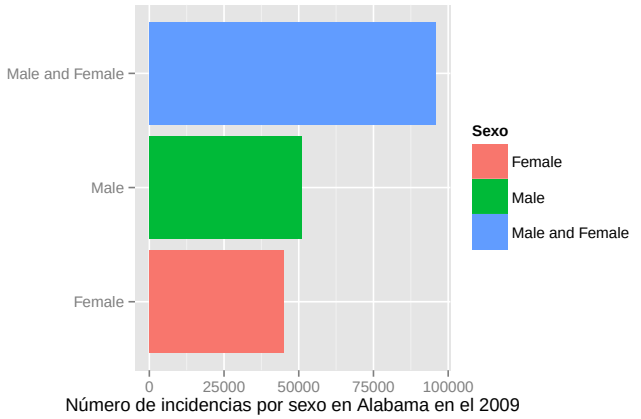

Figura 5. NÚMERO DE INCIDENCIAS POR TIPO DE SEXO EN ALABAMA 2009

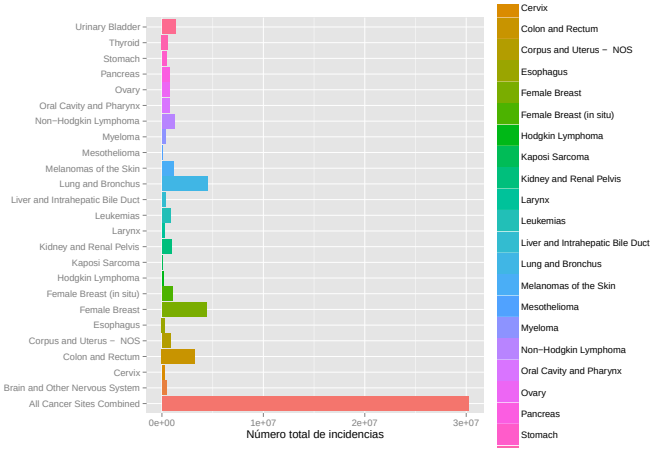

Figura 6. NÚMERO TOTAL DE INCIDENCIAS POR TIPO DE CÁNCER

VI. CONCLUSIONES

Como se ha podido comprobar la tarea de construccin de un data warehouse no es nada trivial y más que la tarea en global, el proceso ETL es la parte de todo el conjunto que más problemas puede presentar y presenta. En este caso no ha sido un proceso demasiado árido y que derivase múltiples problemas debido principalmente a que solo contábamos con una fuente de datos, lo cual facilitaba el trabajo.

Aunque dicha construccin sea una tarea que precisa de un esfuerzo considerable es totalmente compensable. Las aplicaciones y utilidades derivadas de un data warehouse son numerosas, aportándonos un mejor y más rápido análisis de datos complejos, por todo ello, nos puede proporcionar "expectativas" de como será el futuro y poder predecirlo.

En el caso concreto que se expone en este trabajo, el data warehouse presenta múltiples aplicaciones como ya han sido detalladas, pudiéndolas dividir en dos grandes grupos, por un lado el grupo para la visualizacin de datos como podrían ser los mapas de colores u otros gráficos, y por el otro lado las herramientas de predicción o estimación como el clasificador bayesiano.

En resumen, las decisiones que se toman en cualquier ámbito están muy influenciadas si la organización cuenta con un data warehouse, todas estas estrategias y aspectos relevantes que pretenden crear conocimiento sobre el entorno de trabajo, a través del análisis de los datos recibe el nombre de **Buisness Intelligence**, término que nos será familiar en los próximos aos.

REFERENCES

[1] (2014, Febrero) Datos y cifras sobre el cáncer, OMS. [Online]. Available: http://www.who.int/cancer/about/facts/es/