Análisis y creación de un sistema informático de soporte a entidades sanitarias para la detección precoz de brotes de gripe en épocas de máximo riesgo.

Sergio Hernández Gasó
Master Universitario de Ingeniería informática

Consultor: Felipe Geva Urbano.

10/06/2014

FICHA DEL TRABAJO

Título del trabajo:	**Análisis y creación de un sistema informático de soporte a entidades sanitarias para la detección precoz de brotes de gripe en épocas de máximo riesgo.**
Nombre del autor:	*Sergio Hernández Gasó*
Nombre del consultor:	**Felipe Geva Urbano.**
Fecha de entrega (mm/aaaa):	*06/2014*
Área del Trabajo Final:	*Bussines Intelligence*
Titulación:	Master Universitario de Ingeniería informática

Resumen del trabajo (máximo 250 palabras):

Hoy en día, las posibilidades del Big Data son incontables. Existe gran cantidad de información generada por la población general y disponible de forma pública.

El reto consiste en poder trabajar con esta información y extraer conclusiones útiles y que generen valor.

En este proyecto, queremos analizar en el tiempo el interés general de la población respecto a una enfermedad común como la gripe, y poder relacionarlos con brotes de gripe existentes en el pasado, para de esta manera, poder extrapolar y predecir futuros brotes.

Esta información, en manos de las autoridades sanitarias, puede ser de gran ayuda para poder prevenir picos de solicitudes en los servicios de urgencias, anticipándose para gestionar de manera más eficaz los recursos disponibles, consiguiendo, de esta manera, un mejor servicio a la población en general.

De esta manera, son los propios usuarios los que, sin saberlo, posibilitan una mayor y mejor respuesta en los servicios sanitarios mediante la información que ellos mismos distribuyen libremente, consiguiéndose de esta manera valiosos beneficios para la población general.

Abstract (in English, 250 words or less):

Nowadays, the possibilities of Bid Data are endless . There is a great amount of information generated by the general population and available publicly.

The challenge is to work with this information and draw useful conclusions to generate value.

In this project, we analyze over time the general interest of the population on a common illness like the flu , and relate it with existing flu outbreaks in the past. In this way , we could to extrapolate and predict future flu outbreaks.

This information, in the hands of the health authorities, can be of great help to prevent spikes in requests emergency services in anticipation to more efficiently manage available resources, obtaining in this way , a way to better serve the population in general.

Thus , the users themselves are who unwittingly enable a better response in health services through information that they distribute freely, achieving valuable benefits for the general population .

Palabras clave (entre 4 y 8):

Gripe, Big Data, Pentaho, Redes sociales, Análisis predictivo.

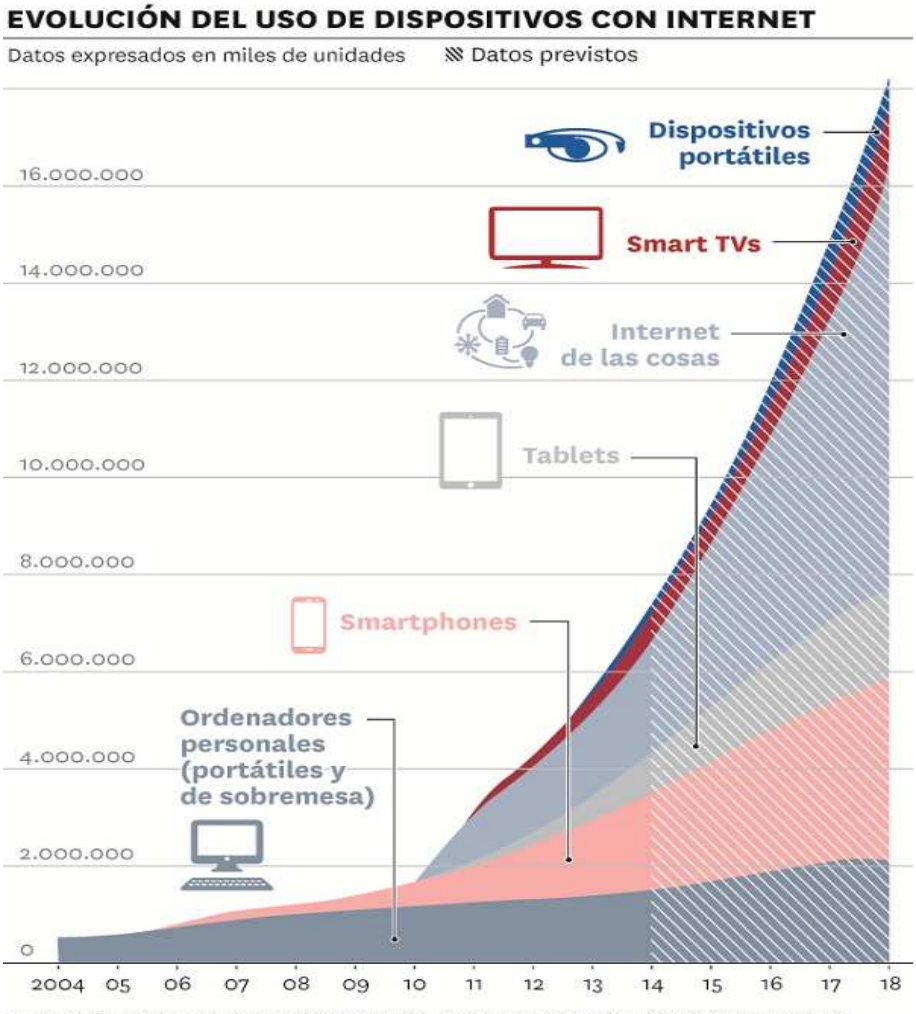
Tabla de contenido

1 Introducción:

1.1 Contexto:

La llegada y popularización de las redes sociales ha puesto a disposición de la sociedad ingentes cantidades de información que son ofrecidas libremente por los usuarios de dichas redes.

El llamado “Big Data” crece exponencialmente año tras año, y esta tendencia no tiene visos de cambiar. Diariamente se postean más de 300 millones de fotos de Facebook, y se postean más de 12 terabits de tuits[1]. Además, recientemente se está popularizando el “internet de las cosas”, donde los objetos más cotidianos están provistos de sensores que arrojan nueva información sensible de ser tratada (zapatillas, neveras, bombillas...).



Fuente: http://www.20minutos.es/noticia/2089892/0/domotica/hogar/internet/

[1] ¿Tuits o tweets? Oficialmente para la RAE, el término es “tuit”, aunque pueda parecer lo contrario. En este texto utilizaremos tuit en todos los casos.
http://espaciosblog.com/tuitweet.html

Todo esta cantidad de información se ve propiciada por el auge de los dispositivos conectados a internet, hasta hace poco territorio exclusivo de los Pcs, pero recientemente con un gran auge de dispositivos móviles y objetos cotidianos
Como muchos autores señalan, la era de la informática está originando la era de la información, y ésta está viniendo a pasos agigantados y de manera exponencial.
Es por ello que es muy interesante ver la manera de poder procesar y utilizar esta ingente cantidad de información en nuestro propio beneficio, ya que disponemos de montañas de información que son inútiles sin sistemas que nos permitan procesarlos y extraer conclusiones.
Y es necesario que estos sistemas estén automatizados porque es virtualmente imposible que podamos procesar toda esta información sin herramientas adecuadas.

1.2 Planteamiento del problema.

La gripe es una enfermedad infecciosa que provoca una incapacitación temporal que puede ser severa en algunos casos, y que puede dar lugar a un uso superior a lo normal en las unidades de urgencias de los hospitales y los centros de salud.

Esta enfermedad no afecta a toda la población de manera uniforme, sino que a lo largo de la época invernal pueden existir uno o varios repuntes que provocan una avalancha en los servicios médicos, lo cual se traduce en un serio detrimento de la calidad y rapidez de la atención médica recibida.
Estos brotes podrían ser tratados con mucha mayor efectividad si se supiera de antemano cuando comienza un nuevo brote, para poder asignar los recursos sanitarios precisos en las áreas donde se prevé una mayor necesidad de ellos.
Se conoce que la población en general posee una cierta tendencia a postear en redes sociales aquellos eventos que le acontecen de manera inmediata, generando comentarios que nos puedan dar pistas sobre su estado de salud, por lo que, en base a dicha información posteada de manera libre, se plantea si es posible extraerla y utilizarla como un sistema de apoyo a decisiones, donde los propios usuarios serán los generadores de la información que nos puede ayudar a distinguir el inicio de un nuevo brote.

Por tanto, y siguiendo la propuesta existente en
http://cv.uoc.edu/app/phpBB3/viewtopic.php?f=7122&t=48226&sid=09407d0d8a18bae399e7a441d484e2f8
, nos planteamos realizar un sistema de predicción de brotes gripales basado en comentarios realizados en redes sociales o consultas en los buscadores.
Ello es posible porque algunas redes permiten realizar consultas sobre la información generada por sus usuarios. Es el caso de twitter, que permite realizar búsquedas por hashtags, pudiendo realizar filtros para los resultados como localización y fecha. Además, hay buscadores como google que ofrecen

de manera gratuita y pública, información puntual y filtrada por países donde nos indica el número de búsquedas realizadas relacionadas con la gripe. Será el objetivo de este proyecto determinar la de qué manera podemos extraer y almacenar los datos existentes en internet para poder trabajar con ellos en la consecución de los objetivos.

Mediante dichos datos, y las bases de datos de los hospitales públicos, podemos crear un sistema BI que permita trabajar relacionando los datos de ambas fuentes, para realizar una correlación que nos permita extraer conclusiones y prever posibles repuntes en infecciones tan comunes, pero al mismo tiempo tan incapacitantes, como la gripe.
No es necesario remarcar lo útil que puede resultar un sistema de dichas características para el sector sanitario, puesto que permite balancear recursos y aplicarlos en aquellos puntos donde vayan a ser necesarios de una manera proactiva, en lugar de reactiva como hasta ahora.
Ello posibilita una mayor racionalización de recursos, una mejor atención al cliente, y por tanto, una eficacia mayor en los sistemas de salud pública, lo cual redunda no tan sólo en dicho sistema, sino en el conjunto de la sociedad, si tenemos en cuenta la mejor cobertura de la población frente a estas infecciones y la mejor atención recibida.

Para la realización del proyecto, se identifican varias fases en la tarea a realizar:

- Obtener información de las redes sociales y/o buscadores
- Cargar la información obtenida en una base de datos apropiada.
- Crear un cuadro de mandos o informes para que los actores implicados puedan obtener la información en los momentos más necesarios y conocer con antelación cualquier posible brote.

Adicionalmente, la aplicación debe ser lo más modular posible, para permitir la ampliación del proyecto a nuevas redes y nuevas búsquedas de información relacionada.
No en vano, una vez que la infraestructura esté instalada y validada, no es muy difícil imaginar las posibilidades que ofrecen las ingentes cantidades de información de las redes sociales en el objetivo de conocer las inquietudes y necesidades de la población en general, en temas sanitarios tan comunes como información sexual, drogadicción, infecciones, tabaquismo o cualquier otro punto que sea susceptible de ser analizado y evaluado.

1.3 Requisitos y objetivos:

El proyecto se podrá considerar que ha tenido éxito cuando se consigan los siguientes objetivos:
-Que posea un sistema o procedimiento capaz de obtener y cargar la información necesaria de la red.
-Que sea capaz de relacionar entre la información que se genera en la red y la histórica existente en las bases de datos de los centros de salud.
-Que se generen conclusiones capaces de predecir cuándo se va a producir un hito de interés para los actores, en este caso, un brote de gripe.
-Que la aplicación sea capaz de comunicar esta información de una manera eficaz a los destinatarios apropiados.
El proyecto se podrá considerar que ha tenido éxito cuando se consigan los objetivos descritos con unas características técnicas y de usabilidad aceptables.

Adicionalmente, existen otros objetivos secundarios que, aun no siendo críticos, nos aumentaría la calidad del proyecto y la satisfacción de los actores:
-El diseño debe ser modular, escalable y ampliable con facilidad.
Se debe tener en cuenta la posibilidad de aumentar la funcionalidad de la aplicación conforme se detecten nuevas necesidades.
-La presentación debe ser clara y concisa. Hemos de tener en cuenta que los destinatarios no tienen por qué ser grandes conocedores del funcionamiento de un BI, o expertos en sistemas informáticos. Es importante que la usabilidad sea la mejor posible, presentando los contenidos de manera efectiva.
-Es deseable utilizar soluciones open source para el desarrollo y utilización de la plataforma. Ello posibilita que se genere una solución de bajo coste y mantenimiento, y que pueda ser utilizada por el mayor número de actores posible.

1.4 Enfoque y método a seguir.

Se plantea un enfoque tradicional de desarrollo en cascada para la consecución del proyecto.

Según dicho enfoque, realizaremos en una primera fase una evaluación detallada de los requisitos en base a la propuesta inicial, una definición exhaustiva de los requisitos, y una planificación de las tareas a realizar.

Una vez terminada dicha fase preliminar, pasaremos al análisis de la aplicación a realizar, refinando los objetivos a un nivel más técnico, concretando los criterios y las restricciones del proyecto, y evaluando el software más adecuado en base a los objetivos planteados.
Es en esta fase también cuando se debe evaluar la viabilidad técnica de las soluciones, y decidir si el proyecto es viable o no.
En el caso de que no fuera viable por alguno de los motivos, deberíamos volver a la fase anterior para reevaluar los requisitos, o bien abandonarlo.

Si se ha llegado a la conclusión de que el proyecto es viable, pasaremos a la fase de diseño, donde definiremos las distintas fases de la solución, concretaremos el diseño de la base de datos, y esbozaremos las transformaciones de datos necesarias, que debe abarcar desde la recogida de datos iniciales hasta su inclusión con la estructura de base de datos final, listo para su uso en el sistema BI.
Además, se debe comenzar a diseñar los informes y el cuadro de mandos que queramos implementar.

Tras todas estas fases, comenzaremos la implementación, donde crearemos los procesos o procedimientos de carga, las transformaciones ETL, y los informes y cuadros de mando.
En cada una de las fases, realizaremos unas pruebas unitarias para verificar la bondad de la codificación.

Por último, terminaremos con una fase de validación, donde realizaremos pruebas de integración de toda la solución al completo, y realizaremos los test de usabilidad pertinentes.

Al margen de las etapas anteriores, se plantea una fase de seguimiento y documentación, no supeditada a las anteriores, que abarque la totalidad del proyecto, y nos indique en cada momento el estado del mismo.

Requisitos	Evaluación de requisitos en base a la propuesta inicial del proyecto. Definir los objetivos del proyecto. Realizar la planificación de las tareas
Análisis	Refinar los objetivos para obtener un mayor detalle. Concretar los criterios de aceptación y las restricciones. Evaluar el software necesario. Verificar la viabilidad técnica de las soluciones propuestas
Diseño	Definir las diferentes fases de la solución. Concretar el diseño de la base de datos a utilizar. Diseñar las transformaciones de datos necesarias. Postular los informes necesarios Esbozar el cuadro de mandos a realizar.
Implementación	Creación de los procesos/procedimientos de carga. Creación de informes y cuadro de mandos. Realización de las transformaciones ETL requeridas. Pruebas de cada una de las fases.
Verificación	Pruebas de integración del proyecto. Test de usabilidad.

Todo el proyecto	Creación y mantenimiento de la documentación asociada.

1.5 Planificación del trabajo

Tareas a realizar

En un primer momento podemos identificar las primeras tareas a realizar, así como el tiempo disponible para las mismas.
Es imprescindible remarcar que es una primera estimación, y que conforme avance el proyecto, es de vital importancia realizar el seguimiento de las mismas, para controlar posibles desviaciones y retrasos que puedan suceder, y realizar las acciones correctivas pertinentes para que los plazos no se retrasen y el plan de trabajo siga vigente.

Las tareas identificadas son las siguientes:

- Formación en recuperación de datos de las redes
- Evaluar soluciones software para realizar el desarrollo y elegir la mejor opción.
- Formación en dicha herramienta.
- Análisis y diseño del sistema.
- Creación de la Base de datos.
- Creación del proceso de carga
- Validación y pruebas (carga)
- Creación de informes
- Creación del cuadro de mandos.
- Evaluación usabilidad del cuadro de mandos
- Validación y pruebas (informes y cuadro de mandos)
- Integración de los procesos
- Pruebas integradas de la aplicación
- Implementación
- Memoria preliminar y documentación
- Revisión y ampliación documentación.

Hitos

10/03/2014 – PAC1- Validación del plan de trabajo y comienzo.
14/04/2014 – PAC2 - Análisis y diseño finalizados.
16/05/2014 – PAC3 - Implementación de la aplicación. Memoria preliminar
10/06/2014 – Entrega – Documentación final.

Distribución de horas y tareas

Tras haber identificado las tareas a realizar, y los hitos a conseguir, pasamos a realizar una primera evaluación de costes y dependencias que nos permita posteriormente realizar una planificación.

Ref	Tarea	Duración estimada	Precedente	Fecha hito.
1	Formación: Recuperación de datos de la red	8 horas		
2	Evaluación de una solución software para crear el sistema BI	4 horas		
3	Formación en la herramienta elegida.	20 horas	2	
4	Análisis y diseño del sistema	20 horas		14/04/2014
5	Creación de la Base de datos.	8 horas	3,4	
6	Creación del proceso de carga	25 horas	1,4	
7	Validación y pruebas (carga)	4 horas	6	
8	Creación de informes	20 horas	4	
9	Creación del cuadro de mandos.	20 horas	4	
10	Evaluación usabilidad del cuadro de mandos	4 horas	9	
11	Validación y pruebas (informes y cuadro de mandos)	8 horas	8,9	
12	Integración de los procesos	8 horas	7,11	
13	Pruebas integradas de la aplicación	8 horas	12	
14	Implementación	8 horas	13	
15	Memoria preliminar y documentación	15 horas		16/05/2014
16	Revisión y ampliación documentación.	30 horas	15	10/06/2014

Estimación: 210 horas.

Planificación

Una vez identificadas y valoradas las tareas, vamos a realizar la planificación del proyecto con los siguientes datos:
-Se considerará una jornada laboral de 4 horas dedicadas en exclusiva a esta asignatura.
-La semana laboral será de lunes a viernes, sin utilizar sábados ni domingos. No obstante, no se van a tener en cuenta días festivos.
-La fecha de inicio de proyecto se establece una vez aprobada la PAC1, donde se finaliza el presente documento y se marca como estable.

Con dichos datos, se crea la planificación del proyecto (página siguiente)

Al desplegar sobre el papel las tareas podemos observar los siguientes puntos:
-El primer plazo de entrega posee una holgura de tiempo considerable. Ello es muy importante, ya que es las primeras tareas (sobre todo las de formación), pueden tener una amplia variabilidad.

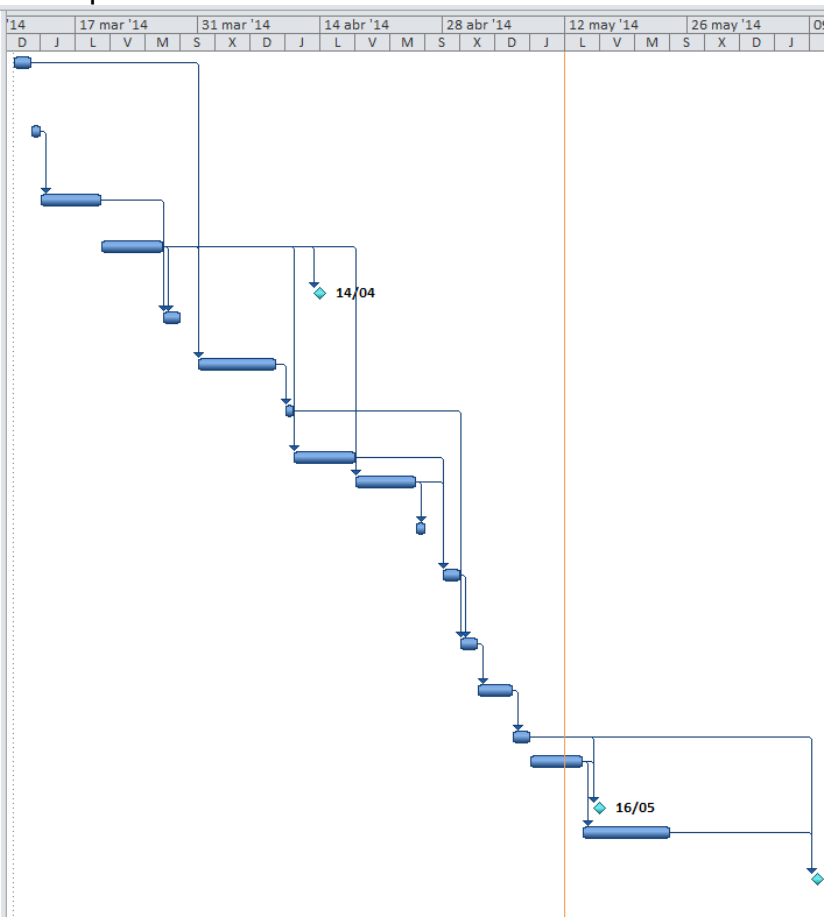
-El segundo plazo es bastante ajustado. Será necesario revisar muy de cerca el desarrollo de las tareas y actuar con prontitud ante cualquier posible desviación.

-El tercer plazo también posee algo de holgura. Ello nos permitirá corregir cualquier tarea que pueda estar incorrecta o no demasiado fina, antes de la entrega final.

Nombre de tarea	Trabajo	Comienzo	Fin	Predecesoras
Formación: Recuperación de datos de la red.	8 horas	lun 10/03/14	mar 11/03/14	
Evaluación de una solución software para crear el sistema BI	4 horas	mié 12/03/14	mié 12/03/14	
Formación en la herramienta elegida.	20 horas	jue 13/03/14	mié 19/03/14	2
Análisis y diseño del sistema	20 horas	jue 20/03/14	mié 26/03/14	
HITO : PAC 2	0 horas	lun 14/04/14	lun 14/04/14	4
Creación de la Base de datos.	8 horas	jue 27/03/14	vie 28/03/14	3;4
Creación del proceso de carga	25 horas	lun 31/03/14	mar 08/04/14	1;4
Validación y pruebas (carga)	4 horas	jue 10/04/14	jue 10/04/14	7
Creación de informes	20 horas	vie 11/04/14	jue 17/04/14	4
Creación del cuadro de mandos.	20 horas	vie 18/04/14	jue 24/04/14	4
Evaluación usabilidad del cuadro de mandos	4 horas	vie 25/04/14	vie 25/04/14	10
Validación y pruebas (informes y cuadro de mandos)	8 horas	lun 28/04/14	mar 29/04/14	9;10
Integración de los procesos	8 horas	mié 30/04/14	jue 01/05/14	8;12
Pruebas integradas de la aplicación	8 horas	vie 02/05/14	lun 05/05/14	13
Implementación	8 horas	mar 06/05/14	mié 07/05/14	14
Memoria preliminar y documentación	15 horas	jue 08/05/14	mar 13/05/14	
HITO : PAC 3	0 horas	vie 16/05/14	vie 16/05/14	15;16
Revisión y ampliación documentación.	30 horas	mar 13/05/14	vie 23/05/14	16
HITO : ENTREGA	0 horas	mar 10/06/14	mar 10/06/14	18;15

Utilizamos el programa Ms Proyect 2010 para colocar las tareas en un diagrama de Gant.
Ello nos permitirá advertir, de forma más visual, cuáles son las tareas más relevantes, y situarlas correctamente respecto a los hitos marcados

	Nombre de tarea	Trabajo	Comienzo	Fin	Prede
1	Formación: Recuperación de datos de redes.	8 horas	lun 10/03/14	mar 11/03/14	
2	Evaluación de una solución software para crear el sistema BI	4 horas	mié 12/03/14	mié 12/03/14	
3	Formación en la herramienta elegida.	20 horas	jue 13/03/14	mié 19/03/14	2
4	Análisis y diseño del sistema	20 horas	jue 20/03/14	mié 26/03/14	
5	**PAC 2**	0 horas	lun 14/04/14	lun 14/04/14	4
6	Creación de la Base de datos.	8 horas	jue 27/03/14	vie 28/03/14	3;4
7	Creación del proceso de carga	25 horas	lun 31/03/14	mar 08/04/14	1;4
8	Validación y pruebas (carga)	4 horas	jue 10/04/14	jue 10/04/14	7
9	Creación de informes	20 horas	vie 11/04/14	jue 17/04/14	4
10	Creación del cuadro de mandos.	20 horas	vie 18/04/14	jue 24/04/14	4
11	Evaluación usabilidad del cuadro de mandos	4 horas	vie 25/04/14	vie 25/04/14	10
12	Validación y pruebas (informes y cuadro de mandos)	8 horas	lun 28/04/14	mar 29/04/14	9;10
13	Integración de los procesos	8 horas	mié 30/04/14	jue 01/05/14	8;12
14	Pruebas integradas de la aplicación	8 horas	vie 02/05/14	lun 05/05/14	13
15	Implementación	8 horas	mar 06/05/14	mié 07/05/14	14
16	Memoria preliminar y documentación	15 horas	jue 08/05/14	mar 13/05/14	
17	**PAC 3**	0 horas	vie 16/05/14	vie 16/05/14	15;16
18	Revisión y ampliación documentación.	30 horas	mar 13/05/14	vie 23/05/14	16
19	***Entrega***	0 horas	mar 10/06/14	mar 10/06/14	18;15



1.6 Sumario de productos obtenidos.

A la finalización de este proyecto, se deben tener listos los siguientes entregables:
-El presente documento, con la documentación completa del proyecto.
Esta documentación debe incluir:
- -Esquema de las bases de datos
- -Pruebas realizadas
- -Evaluación de la usabilidad.

-Procesos realizados para la carga de datos.
-Procesos ETL para la transformación de los datos (exportados a XML)
-Informes y cuadros de mando generados.
-Video explicativo del funcionamiento del proyecto.
-Vídeo con las pruebas de usabilidad realizadas.

1.7 Breve descripción de otros capítulos de la memoria.

Comenzaremos en el capítulo 2: Análisis del proyecto. , con un análisis en profundidad sobre los requerimientos del proyecto, poniendo en claro sus objetivos y restricciones.

Una vez tengamos los objetivos claros, el capítulo 3:Elección de herramientas a utilizar. versará sobre las distintas alternativas para su implementación y los criterios que se han seguido para elegir una u otra herramienta.

Al tener los objetivos claros y las herramientas elegidas, en el capítulo 4 : Diseño realizaremos un diseño completo de cada una de las fases.

En el capítulo 5: Validez y precisión de los datos., intentaremos establecer unos KPI que nos permitan definir exhaustivamente el funcionamiento interno del cuadro de mandos, así como los valores finales que se mostrarán al usuario.

Una vez tenemos los objetivos, el diseño, las herramientas y los KPI, pasaremos a la fase de implementación, comentada en el capítulo 6: Implementación.,

La fase de pruebas se diseña en el capítulo 7: Pruebas realizadas, pero se documenta de manera completa en el Anexo 1: Pruebas.

Los test de usabilidad se diseñan en el capítulo 8: Pruebas de usabilidad. Los resultados se indican en el Anexo 2: Pruebas de usabilidad.

Por último, reseñar los capítulos 9: Inclusión de datos de redes sociales., donde comentamos la posibilidad de ampliar el proyecto con un mayor número de redes sociales, o el 10: Conclusiones, donde comentamos las conclusiones extraídas durante el desarrollo de este proyecto.

2 Análisis del proyecto.

2.1 Objetivos específicos:

En base a la información anterior, podemos establecer los siguientes requisitos para la correcta realización del proyecto:

-El sistema debe ser capaz de recoger información de la red. Se plantean dos alternativas. La primera de las cueles es obtener información recogida de la red Twitter, por las siguientes características:

-La información se generada en forma de texto, lo cual hace viable su análisis.

-Existe una API generada por la red que permite la búsqueda y descarga de información posteada por cualquier usuario de la red, pudiéndose filtrar además de acuerdo a unos requisitos de contenido, fecha de generación y posición geográfica que nos permita seleccionar aquellos datos más interesantes.

No obstante, y tras unos estudios preliminares, advertimos que esta vía posee una importante limitación, y es la imposibilidad de poder recuperar información con una antigüedad superior a una semana. Ello imposibilita el contrastar valores históricos hasta que no recojamos por nuestra cuenta una cantidad de ellos suficiente para su análisis.

La segunda opción es obtener los datos desde la publicación que realiza google sobre las consultas realizadas por sus usuarios sobre la gripe.

Esta información también se obtiene en formato texto, y nos permite filtrarla por región, y recuperar un histórico de datos muy amplio, de varios años.

Se plantea la posibilidad de obtener la principal fuente de dados desde google, y añadir los datos que se puedan incorporar de twitter como una segunda vía que aporte información extra.

-La información obtenida de esta fuente debe almacenarse en una base de datos específica de carga.

-Se proveerá una serie de datos iniciales por parte del cliente que proporcione los datos a contrastar y de referencia para poder trabajar. Dichos datos deben filtrarse, limpiarse y cargarse en una base de datos.

-Se debe crear un almacén de datos que soporte el almacenamiento tanto de los datos propios del cliente como de los extraídos en las redes sociales.

-Asimismo, deben existir unos procesos ETL que alimenten el almacén de datos a partir de las bases de datos de inicio y carga

-Deben existir herramientas que permitan tratar y relacionar los datos existentes en el almacén para poder extraer conclusiones.

-Se definirán y crearán una serie de indicadores y valores de referencia que nos permitan predecir con una exactitud aceptable en qué momento podemos ser susceptibles de sufrir un brote de gripe.

-Se deben desarrollar una serie de funcionalidades que calculen de manera automática los indicadores anteriores, y por tanto, la probabilidad de que se

genere un brote de gripe, e informar de ello al cliente de la manera más adecuada posible. Para ello se propone la creación de un cuadro de mandos con la información pertinente en pantalla y/o la generación y envío de informes a aquellas personas que estén interesadas o sea de especial relevancia que permanezcan informadas de este evento.

Factores críticos para el éxito del proyecto:
-Se debe demostrar la fiabilidad del sistema. Para ello, habrá que identificar los brotes de gripe en la base de datos de ICS de años anteriores y relacionarlo con los datos obtenidos de la web.
-El proceso debe estar terminado y validado antes del 10/06/2014

2.2 Requisitos del proyecto:

-Ser capaz de cargar correctamente la información de la red en base a unos parámetros definidos. La velocidad del proceso no es relevante, pero debe realizarse en un tiempo asumible.
-Integración de los datos recogidos con la base de datos ya existente. Posibilidad de actualizar la información proveniente de una u otra fuente de manera rápida y sencilla.
-Capacidad para relacionar la información del ICS con la recogida de fuentes externas, de manera visual.
-Creación, soporte y actualización en tiempo real de los indicadores que nos puedan indicar un posible brote de gripe. Demostración de la validez de dichos indicadores.
-Generación de un cuadro de control que nos permita advertir de rápidamente del estado de dichos indicadores. Opcionalmente se puede añadir el envío de informes a las entidades interesadas que no posean acceso a dicho cuadro.

2.3 Criterios de aceptación del proyecto:

-Todos los requisitos obligatorios deben ser cumplidos.
-No deben existir errores graves en la solución aportada.
-La validez y fiabilidad de los indicadores debe ser demostrada.
-Se debe aportar la documentación necesaria para el correcto uso de la herramienta.
-El diseño debe ser modular para incluir futuras nuevas fuentes de datos.

2.4 Restricciones al proyecto.

No entra dentro del alcance del proyecto:
-La formación a los usuarios.
-La recuperación de datos de fuentes distintas a las planteadas.

3 Elección de herramientas a utilizar.

Una vez establecidos los requisitos, pasaremos a analizar las herramientas que debemos utilizar, así como las restricciones que nos puede acarrear el seleccionarlas.
Según los requerimientos, necesitaremos al menos dos herramientas:

-Una herramienta o procedimiento para el proceso de carga. Con esta herramienta debemos obtener los datos de internet necesarios y cargarlos en una BD relacional. Adicionalmente, sería interesante tener cargados también los datos suministrados por el ICS en la misma base de datos.

-Otra herramienta de análisis y presentación de los datos (BI). Esta herramienta debe permitirnos relacionar los datos y generar una representación gráfica y amigable de los mismos, preferentemente con un cuadro de mandos, aunque no se descartaría generar informes puntuales, dando preferencia a la primera opción.
-Dependiendo de las capacidades de las herramientas anteriores, es probable que se requiera una tercera, para realizar procesos ETL entre una y otra. Es muy probable que las suites de BI existentes incluyan también esta funcionalidad, pero hay que tenerlo en cuenta.

3.1 Herramientas de carga:

Nos planteamos la fuente de datos más apropiada que para el proyecto actual. Evaluamos la posibilidad de obtener los datos de la red social twitter.
Para ello, necesitaríamos una herramienta capaz de recuperar los tuits en determinadas épocas y almacenarlos en una base de datos.
Para ello, pensamos que la mejor opción será utilizar la API de twitter, en un programa en java.
Revisamos la API de twitter y buscamos algún wrapper que nos permita utilizarlo con Java, y que esté actualizado con la API 1.1, ya que la API 1.0 ya no está vigente.
Lo más característico de la nueva API es que los programas deben incluir una serie de claves privadas a cada usuario para la utilización de los recursos.

Existen varios wrappers para utilizar la API de twitter con java, entre ellos el HBD (https://github.com/twitter/hbc) y el twitter4j (http://twitter4j.org/en/index.html), ambos actualizados a la nueva API y con relativamente abundante información como para poder hacer una aplicación con garantías.
Como base de datos a esta aplicación, una de las que presentan mayor compatibilidad con otras aplicaciones, es gratuita, y posee un buen rendimiento es MySql (http://www.mysql.com/)

Realizamos una pequeña aplicación con todos estos componentes para comprobar la viabilidad de la aplicación.
Tras realizar algunas pruebas, advertimos que nos es imposible recuperar tuits con una antigüedad superior a un mes.

Revisamos en profundidad el API de twitter en busca de alguna explicación, y la encontramos en https://dev.twitter.com/docs/using-search

"*Please note that now API v1.1 requires that the request must be authenticated, check Authentication & Authorizationdocumentation for more details on how to do it. Also note that the search results at twitter.com may return historical results while the Search API usually only serves tweets from the past week. "*

Por tanto, debemos abandonar esta opción y buscar alternativas.

Existen herramientas alternativas a la API de twitter como por ejemplo http://topsy.com/. Verificamos que sí que se pueden añadir filtros de tiempo, pero la salida está limitada siempre a 100 tuits y no informa del número de tuits existentes en la selección. Existe una opción de social analitics, donde se generan gráficas con tuits a lo largo del tiempo, pero no se puede fijar el rango temporal, establecido en los 7 últimos días.

Existe un wrapper en java para Topsi, (http://code.google.com/p/otter4java/), pero al parecer dejó de actualizarse el 16/12/2012. A partir del 01/08/2013, topsi necesita de una API para funcionar, por lo que esta fuente ya no es válida.

Otras alternativas han sido discontinuadas o son servicios de pago nos ofrecen información desde el momento de su contratación (http://www.tweetarchivist.com/). Todas adolecen del mismo problema: no monitorizan más allá de 30 días.

Parece que actualmente no es sencillo obtener información directamente de twitter con la antigüedad requerida, lo que nos obliga a replantear el proceso de carga.

Una alternativa realmente accesible y viable parece ser el estudio previo que realizó Google sobre este mismo problema (http://www.google.org/flutrends). Google ha estado guardando información sobre las búsquedas realizadas por los usuarios sobre este tema, información que nos permite descargar en forma de búsquedas semanales agrupadas por países y regiones.

Revisamos la información aportada, verificando que los datos aportan la antigüedad necesaria, y que existe una continuidad hasta el momento actual.

Se propone, por tanto, la utilización de estos datos como base para realizar el proyecto.

Como herramientas a utilizar en la carga, crearemos una BD MySql donde cargaremos tanto los datos provenientes del ICS como los datos obtenidos de Google.

El procedimiento que utilizaremos para cargar los datos de google en la base de datos será mediante inserción directa con SQL.

Adicionalmente, se plantea la posibilidad de crear las estructuras de bases de datos necesarias para poder utilizar los datos de twitter que puedan ser obtenidos desde distintas fuentes. Estos datos pueden utilizarse como apoyo para validar y contrastar los resultados obtenidos.

3.2 Herramientas de transformación y presentación.

Debemos elegir una herramienta que nos permita transformar los datos iniciales, analizarlos, y presentarlos de manera ágil y comprensible.

Los factores que utilizaremos en la discriminación del software BI serán:
-A ser posible, que sea gratuito y/o opensource.
La gratuidad de la aplicación nos permitirá que la aplicación sea accesible a un mayor número de receptores, y la característica de opensource nos servirá para apoyar la labor desinteresada de la comunidad de programadores que facilita a la sociedad herramientas de alta calidad al alcance de usuarios y pymes.

-Fácil de usar. Dado que el tiempo de desarrollo es bastante limitado, y que es el propio desarrollador el que debe formarse y resolver todos los problemas que puedan acaecer, es muy importante que la herramienta sea fácil de usar y exista abundante información sobre ella para que el proyecto no sufra problemas inesperados que alarguen el tiempo de desarrollo.

-Que pueda crear cuadros de mando, para poder mostrar la información de la manera más sencilla y comprensible posible.

Analizamos los siguientes programas:

Software comercial con versión gratuita:

-Qlick, edición personal. (http://www.qlik.com)

Qlick ofrece una versión gratuita bastante recortada, con muchas limitaciones para importar/exportar datos, e incapaz de generar cuadro de mandos, tan sólo informes.

-Jaspersoft BI (http://www.jaspersoft.com/es)
La compañía Japerson proporciona varios niveles de su software, incluyendo una versión comunity gratuita.
Dicha versión permite obtener datos de varias fuentes y generar informes en múltiples formatos.
No permite generar cuadro de mandos con esta versión.

Software open source:

-Pentaho. (http://www.pentaho.com/)
Una de las suites más completas, se oferta en versión gratuita (versión community) y de pago. La versión gratuita permite etl, data mining, reportes, cuadro de mandos.
La versión de pago incluye soporte técnico y documentación, pero por lo demás ambas soluciones permiten las mismas funcionalidades.

-Spago BI(http://www.spagobi.org/)
Más que una suite, es una colección de herramientas gratuitas. No existe versión de pago, aunque sí se puede contratar soporte adicional. Capaz de realizar todas las funciones, con diversos motores para cada una.

Parece que, dentro de las versiones gratuitas, el producto con más funcionalidad es el SpagoBI, aunque Pentaho es la suit de referencia del sector. Según la documentación leída, es posible que sea más sencillo de utilizar Pentaho para un programador inexperto, ya que aporta una única solución integrada para todos sus aspectos.
Ambas son herramientas open source, y además con una importante presencia en la red y abundancia de documentación.
También tienen capacidad de generar cuadros de mando, y poseen herramientas ETL propias, por lo que no será necesario utilizar ningún software adicional.

Por tanto, después de este análisis, pensamos que la herramienta a utilizar en esta fase debe ser **Pentaho**.

4 Diseño

Realizaremos un diseño del sistema en tres partes claramente diferenciadas e independientes entre sí. Ello nos asegurará un diseño modular que puede ser atacado por varios equipos de trabajo al mismo tiempo, o bien puede ser actualizado y modificado sin que los cambios se propaguen de manera no controlada al resto de las partes.
Distinguiremos pues tres fases:
-La fase de carga.
-La fase de almacenamiento y procesamiento de datos
-La fase de presentación e informe

4.1 Fase de carga:

Definición de las tablas

En esta fase partimos de datos de varios orígenes diferentes, que debemos combinar y armonizar para poder ser almacenados y tratados en las siguientes fases. Además, las diferencias no radican tan sólo en el formato y origen de los datos, sino en su filosofía, ya que por una parte tenemos datos semanales y por otra impactos diarios.
Atendiendo al volumen de datos de las distintas fuentes, vamos a realizar las siguientes consideraciones:

-Al recuperar datos de google, incorporaremos tan sólo aquellos correspondientes a la región en estudio. Añadir más datos no es de ninguna utilidad para el proyecto, por lo que realizaremos el filtrado en este nivel.

-Para los datos provenientes del ICS, guardaremos toda la información, dejando para futuras fases el filtrado según los valores que consideremos más convenientes. Ello es así porque es más difícil discriminar qué valores son los relevantes, debiendo guiarnos según el área que se ha ocupado de la urgencia/hospitalización.
Observamos que el número de registros a tratar es relativamente pequeño (-64000 registros), por lo que es prudente no filtrar los datos en este paso para evitar eliminar más registros de los necesarios.

-Al recuperar datos de redes sociales, dado el ingente volumen de datos, debemos realizar un filtrado que nos permita cargar tan sólo aquellos datos que nos pudieran parecer relevantes.

Asimismo, y para facilitar una ligera inspección de los datos, guardaremos algunos campos extras que nos permitan analizar la información más fácilmente.

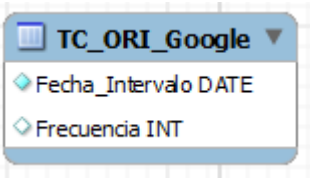
En ningún caso se guardará en ninguna de las tablas información sensible protegida por LOPD

En este diseño, poseemos los siguientes orígenes de datos:

-Datos recogidos de internet. Podemos obtener datos de distintas fuentes, por lo que intentaremos realizar un diseño lo más flexible posible que nos permita incorporar y tratar datos desde distintas fuentes.

Para los datos procedentes de google, advertimos que poseemos la información consistente en búsquedas semanales sobre la gripe. Almacenaremos la siguiente información:

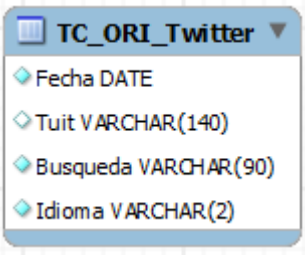
-Semana de análisis. – Formato AAAA/MM/DD
-Número de búsquedas –entero.



Para los datos obtenidos en la red social Twitter. Recuperaremos y almacenaremos los siguientes datos:

-Fecha del tuit. – Formato AAAA/MM/DD
-Texto del tuit. – Máximo 140 caracteres.
-Palabras de búsqueda – criterios por los que se ha recuperado este tuit.
-Idioma – Idioma del tuit

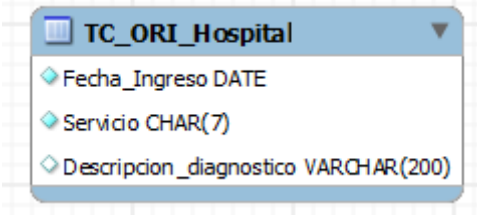
Conforme hemos comentado anteriormente, no se almacenará ningún dato de tipo personal (nombre de usuario, referencia del tuit) para evitar problemas con la LOPD.



-Datos de la ICS

En este caso poseemos dos conjuntos de datos diferenciados: los de hospitalización y los de urgencias. Dado el bajo volumen de datos existentes, no filtraremos en este paso los registros, posponiéndolo hasta la fase de ETL

Para los datos de hospitalización, seleccionaremos
-Fecha de ingreso – Formato AAAA/MM/DD
-Servicio – {Cardiología, Neumología,...}
-Descripción diagnóstico – Máximo 200 caracteres



En el caso de los datos de urgencias, necesitamos los siguientes datos:
-Fecha de entrada.
-Servicio de entrada
-Servicio de alta

La nomenclatura de las tablas se conformará con TC (tabla de carga) + ORI (origen)

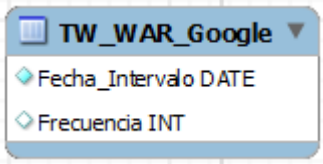
4.2 Fase de Almacenamiento y procesamiento de datos

Esta fase tiene dos tareas principales:
-Crear la estructura de almacenamiento de datos de trabajo (datawarehouse)
-Crear los procesos ETL para cargar la estructura anterior con las bases de datos generadas en el proceso de carga.

Estructura de almacenamiento:
La estructura de datos del datawarehouse constará de una tabla para cada uno de los orígenes de datos. Para optimizar el acceso, dispondremos las tablas en forma de estrella, de manera que los datos relevantes se almacenarán como un dato numérico con correspondencia de su significado en otra tabla.
Ello puede no ser estrictamente necesario en este diseño, donde se espera un número de datos reducido, pero puede ser interesante con visos a escalar el diseño con mayor funcionalidad.

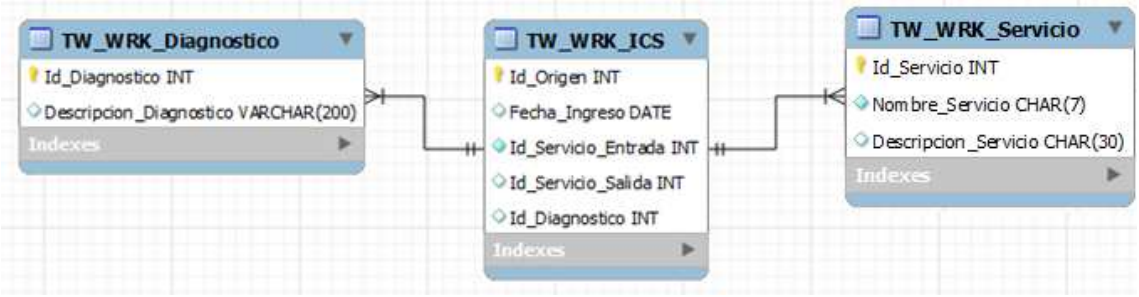
Para los datos provenientes de google, guardaremos los siguientes datos:

-Semana de análisis. – Formato AAAA/MM/DD
-Número de búsquedas –entero.

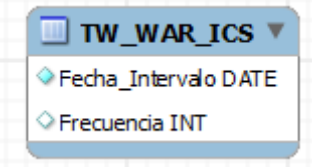


Estos datos son los mismos que ya teníamos en el proceso de carga. No es necesario realizar transformación alguna.

Para los datos provenientes del ICS, guardaremos los siguientes datos:
-Origen: 1 para urgencias y 2 para hospitalización. No es necesario utilizar tablas auxiliares.
-Fecha de ingreso, Formato AAAA/MM/DD
-Servicio de entrada. Para todos los orígenes.
-Servicio de salida. Sólo existentes para urgencias (origen = 1)
-Diagnóstico. Sólo existentes para hospitalización (origen = 2)
Los campos de Servicio y Diagnóstico se extraerán a tablas auxiliares.



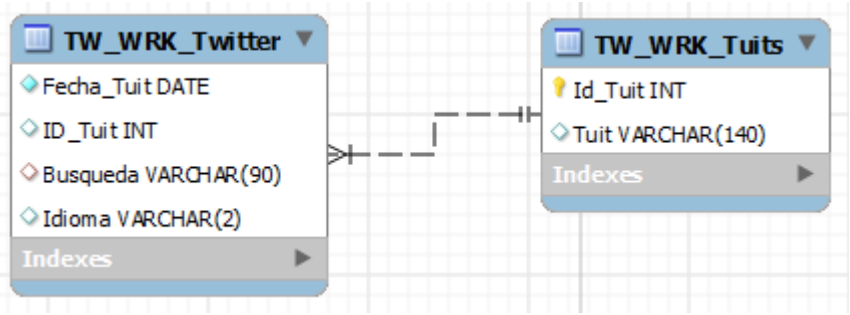
Hay que reseñar que esta tabla es de trabajo (_WRK_)
Dado que los datos que poseemos de google se almacenan como impactos semanales, es interesante generar una tabla final con el mismo criterio.
Por tanto, utilizaremos la tabla de trabajo anterior para buscar y refinar los criterios de filtrado, que posteriormente volcaremos a la siguiente:



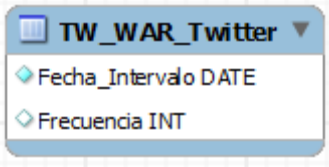
Para los datos provenientes de las redes sociales, guardaremos los siguientes datos:

-Fecha del tuit. – Formato AAAA/MM/DD
-Texto del tuit. – Máximo 140 caracteres.
-Palabras de búsqueda – criterios por los que se ha recuperado este tuit.
-Localización – Si la hubiera.
-Idioma

El texto del tuit se extrae a otra tabla para aumentar el rendimiento.



Dado que los datos que poseemos de google se almacenan como impactos semanales, es interesante generar una tabla final con el mismo criterio.
Por tanto, utilizaremos la tabla de trabajo anterior para buscar y refinar los criterios de filtrado, que posteriormente volcaremos a la siguiente:



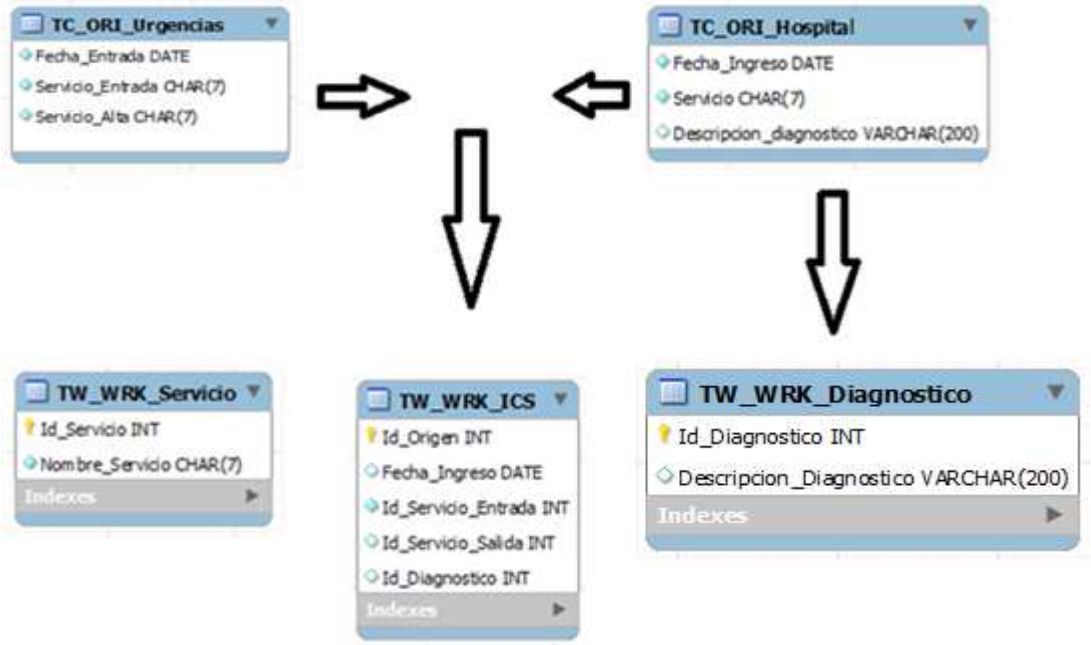
La nomenclatura de las tablas se ha realizado conformará con TW (tabla Warehouse) + WRK (trabajo) o WAR (Tabla final warehouse)

Procesos ETL de carga

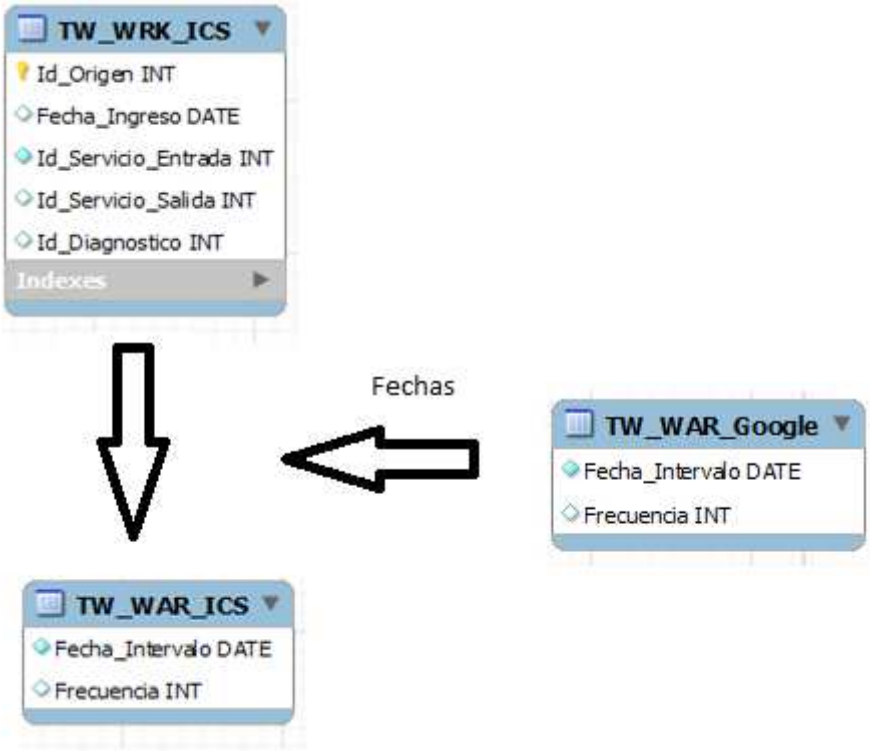
Identificamos varios procesos de carga que hay que realizar para utilizar los datos.

1.- Carga de los datos de google:
Es el proceso más sencillo.
Consistirá en volcar directamente la tabla TC_ORI_Google en TW_WAR_Google.
De esta manera, tras revisar y validar los datos existentes en la tabla de carga, informaremos la tabla destino con los datos deseados.
Tras realizar el análisis de los datos en la fase de diseño de informes, es probable que se detecte alguna anomalía en los datos que pueda falsear los valores mostrados en los informes o en el cuadro de mandos.
Por tanto, realizamos este proceso para poder descartar los datos que nos puedan interesar de la tabla de origen, sin perder esta información.

2.-Carga de los datos del ICS.
Este proceso consta de dos partes.
Por una parte, unificaremos los datos de origen en una sola tabla.



De esta manera, cargaremos la tabla TW_WRK_ICS sumando los registros de las tablas de origen TC_ORI_Urgencias y TC_ORI_Hospital.
En un principio, se utilizará como discriminante el servicio de entrada, empezando por los datos del departamento de Neumología y Medicina general. El posterior análisis de los datos puede sugerir modificar esta selección inicial.
La tabla TW_WRK_Servicio se cargará manualmente con SQL. Podríamos extraer esta información de las mismas tablas de datos, pero ello podría provocar que algún departamento no se incorporara en sucesivas cargas de datos. Por ello, se creará un SQL con los datos de todos los departamentos existentes y se cargará de manera manual.
Por otra parte, se detectan 3 departamentos que sólo aparecen en "Servicio_Alta" de los datos de urgencias. No poseemos la descripción de dichos departamentos (ANCSVGT, APASVGT, IMESVGT), por lo que los cargaremos sin descripción.
La tabla TW_WRK_Diagnostico se cargará en el proceso ETL desde la tabla TC_ORI_Hospital.
Posteriormente, se agruparán los datos por semanas, de la siguiente manera:

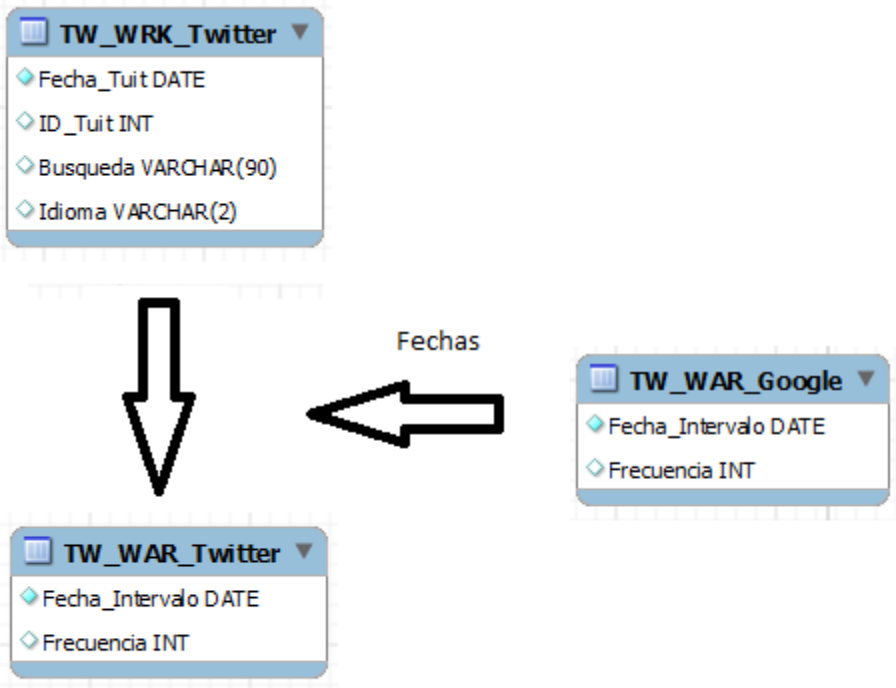


Una vez terminado este proceso, se tendrán cargadas las tablas de datos finales TW_WAR_ICS y TW_WAR_Google

3.-Carga de los datos de twitter.

Se realizará de manera análoga al caso anterior. Partimos de la tabla TW_WRK_Twitter, la cual nos informa de impactos diarios. Realizamos un proceso de transformación para que posea el mismo formato que los datos de google, los cuales indican impactos semanales.

Para ello, realizaremos una selección en base a las fechas ordenadas existentes en la tabla de google, para que no exista ningún error en la transcripción y poder utilizar en todo momento datos nuevos que se puedan incorporar posteriormente.



4.3 Diseño de informes.

Una vez tenemos los datos correctamente cargados, definimos los siguientes informes que nos ayudarán a evaluar la calidad y adecuación de los datos existentes en la base de datos.

Las principales funciones de los informes en este proyecto serán las siguientes:

1.- Revisar la información que se está teniendo en consideración.
Es fácil que los datos puedan quedar falseados si la selección no es correcta, por lo que se plantea realizar una serie de informes que nos aporten información sobre dichos datos.
2.- Obtener una información histórica que demuestre la validez del proyecto.

Identificamos los siguientes informes que pueden aportar claridad y visibilidad a los datos cargados:

-Informe sobre diagnósticos por fecha y servicio.
El propósito de este informe es verificar si los servicios seleccionados son los adecuados para el seguimiento de la epidemia de gripe.
Por ello, vamos a mostrar, agrupados por fecha y servicio, los diagnósticos de cada uno de los casos que se han tenido en cuenta para el estudio.
La revisión de estos diagnósticos nos puede indicar si es necesario eliminar los datos de algún servicio para el propósito de intentar anticiparnos a un repunte de casos gripales.

-Frecuencia de atenciones por servicio y fecha.

Se diseña este informa con la vista puesta en revisar aquellas frecuencias en los servicios que pudiesen servir como base de estudio para este proyecto o cualquier otro.
La utilidad principal de este informe es advertir picos de frecuencia elevados para poder relacionarlos con acontecimientos o situaciones, lo cual nos permitiría establecer una pauta de utilización que podemos utilizar en nuestro beneficio.

Los datos a mostrar será la frecuencia absoluta por servicio y período semanal.

Frecuencia comparativa Google-ICS-Twitter.

Se diseña este informe con la mirada puesta en advertir correlaciones entre las distintas fuentes de datos y establecer cuáles de ellas pueden ser adecuadas para la predicción de futuros brotes de gripe.

En base a esta información, podremos discernir qué fuentes de datos son apropiadas para el estudio, y cuales no pueden ser aprovechadas por su escasa correlación o relevancia con el campo de estudio.

El informe se compondrá con una simple gráfica donde se muestren las frecuencias semanales de cada una de las fuentes tomadas en consideración.

En un principio, las fuentes a utilizar son:
-Datos provenientes de google (google flu)

-Datos recogidos de twitter (búsqueda: gripe, lenguaje: es)
-Datos del ICS

4.4 Diseño del cuadro de mandos.

La creación de un cuadro de mandos se considera la cumbre del proyecto, puesto que es la parte más visual para el cliente, y aquella con la que se relacionará más a menudo, junto con los importes.
Para ello, además de los criterios técnicos, se ha de tener muy en cuenta las necesidades específicas del cliente, y realizar un diseño que cumpla las siguientes características:

-**Debe existir la información relevante**. Se debe poder verificar de un vistazo la información más importante relacionada con el propósito del proyecto. El cliente ha de poder verificar no sólo los datos finales que queremos mostrarle, sino debe conocer en todo momento cómo se han conseguido esos datos. Hay que pensar que, su experiencia en el negocio, puede establecer a posteriori relaciones entre los datos que no han sido detectadas en tiempo de diseño, cristalizándose en nuevas mejoras para el proyecto.

-**No debe mostrarse información irrelevante**. Puede parecer sorprendente, pero no hay que caer en la tentación de incluir aquellos datos que realmente no van a aportar valor ninguno al cliente, y simplemente van a distraer su atención. Existen demasiadas soluciones comerciales que muestran cantidades ingentes de información, gráficos y estadísticas en una misma pantalla, y realmente lo único que consiguen es apabullar al usuario. ¿Es necesario recargar sensorialmente al cliente con la única finalidad de intentar aparentar una potencia y eficacia que realmente estamos perdiendo? Por todo ello, hay que evitar mostrar aquellos datos no relacionados.

-**El cuadro debe ser lo más simple e intuitivo posible**. Este axioma estará casi conseguido si se cumplen los criterios anteriores, pero además hay que saber organizar la información de manera intuitiva y eficaz, simplificando la presentación y agrupándola de manera lógica para que pueda ser observada y asimilada en la menor cantidad de tiempo, incluso para usuarios no habituados a los cuadros de mando.

-Se deben conseguir los criterios de utilidad y precisión. Una vez presentada la información relevante, y sólo ésta, de manera simple y eficaz, debemos permitir al usuario un cierto control sobre los datos mostrados, de manera que pueda trabajar con ellos y extraer conclusiones que le aporten valor añadido a su tarea.

Por todo ello, comprendemos que el diseño de un cuadro de mandos es un proceso íntimamente relacionado con el usuario, y que lo ideal es ir trabajando con él codo a codo hasta llegar a un modelo simple, útil y usable, que se corresponda con aquello que el cliente está esperando, y lo reconozca como

una herramienta eficaz para apoyarle en sus funciones habituales. Tan sólo logrando que el usuario final reconozca la herramienta como un proceso propio que ha sido desarrollado con su ayuda, lograremos que el proyecto sea utilizado en el día a día como parte de los recursos disponibles.

Por el contrario, una herramienta creada a espaldas del usuario, sin contar con su colaboración y sugerencias, es, en un gran número de casos, uno de los motivos de fracaso de un proyecto, no a nivel técnico, sino a nivel funcional, ya que es posible que no se desarrollen funcionalidades que el usuario considera necesarias, y por ello llegue a entregarse el proyecto pero no usarse.

En resumidas cuentas, para este diseño lo ideal sería realizar un modelo basado en prototipos, evaluados por el cliente, hasta llegar a un desarrollo final que sea útil, eficaz, y aceptado.
Proponemos para este proceso comenzar por un desarrollo en lápiz y papel, para después comenzar un desarrollo rápido incremental que nos conduzca al producto final.

Por las especiales características de este proyecto (no existe un papel de "cliente" como tal), substituiremos el modelo de prototipos por una prueba de usabilidad sobre el producto final, que nos aportará información sobre la facilidad de uso y comprensión de los datos del cuadro de mandos generado.

5 Validez y precisión de los datos.

Hasta este momento hemos buscado la fiabilidad y precisión del modelo, pero tan importante como que la programación esté bien construida y que sea fiable es que el modelo sobre el que se sustenta sea igual de fiable y preciso.
No sirve de nada dar una solución correcta cuando el enunciado del problema está equivocado, por lo que es del mayor interés obtener y validar de manera fehaciente las consideraciones que se han tenido en cuenta en el desarrollo de este proyecto.

5.1 Selección de los departamentos del ICS sensibles.

La primera de las premisas que debemos validar es si hemos tomado una muestra representativa de los datos, especialmente en los datos del ICS, donde tenemos información de diversa índole, y donde se podrían seguir diferentes interpretaciones que pueden no ser correctas.
De toda la base de datos existente en el ICS, existen casos de hasta 47 diferentes departamentos, y la duda planteada es "¿hemos elegido los departamentos correctos?".

Para ello, nos vamos a apoyar en el informe "Frecuencia de atenciones por servicio y fecha", generado en el apartado de "Creación de informes",
Debemos obtener las frecuencias relativas de cada departamento para verificar si la frecuencia de casos de cada uno puede haber sido modificado con motivo de la epidemia de gripe, que es el objetivo final de este proyecto.
Tomaremos como valor de referencia el departamento de "Medicina de Urgencias", el cual sabemos positivamente que sí está influenciado por la gripe, y, en base a éste, verificaremos el resto de departamentos para ver si su variabilidad con el tiempo está relacionada con el departamento de control.

Revisamos visualmente dicho informe, en busca de valores que nos puedan indicar, a simple vista, qué departamentos pueden estar más influenciados por la gripe.

En una primera inspección, observamos que el departamento de pediatría parece estar relativamente correlacionado con el de cirugía general.
Además, vemos un gran número de atenciones en el departamento de ginecología, aunque la relación con los anteriores no es demasiado clara, además, que, lógicamente, podríamos esperar una mayor afinidad con otro tipo de departamentos, como el de neumología. En este último, podríamos esperar un mayor número de casos en estos meses, pero no lo parece.
¿Cómo concretar qué departamentos están o no relacionados? Sin duda, necesitamos un análisis más exhaustivo que una simple exploración visual,

puesto que, aunque podríamos intuir alguna relación entre algunos departamentos, no existe una relación nítida que nos lo confirme.
Además, en un punto tan importante como la selección de los datos básicos para el proyecto, necesitamos algún procedimiento formal que nos confirme, sin asomo a dudas, que hemos elegido la información correcta.

Por tanto, vamos a realizar un análisis estadístico que nos confirme, sin asomo a dudas, qué departamentos debemos incluir en el estudio.

En un primer lugar, obtenemos la población sobre la que realizar el análisis.
Vamos a seleccionar de los datos de urgencias + hospitalizaciones aquellos departamentos que posean un número de impactos igual o superior al 2'5% del total.
Mediante esta operación, nos aseguramos de que la muestra siga siendo representativa (con un 95% de confianza), mientras que reducimos el número de casos de estudio.

El total de impactos por departamentos es el siguiente:

```
select servicio, sum(frecuencia) from (
SELECT servicio_entrada as servicio, count(*) as frecuencia FROM
mydb.tc_ori_urgencias
group by servicio_entrada
union
select servicio, count(*) as frecuencia  from mydb.tc_ori_hospital
group by servicio ) z
group by servicio
```

Lanzamos la query anterior en la base de datos, obteniendo los siguientes resultados:

Departamento	Ingresos
HEMSVGT	1
HIVUFGT	1
UOMUFGT	1
UPMUFGT	1
ONCSVGT	3
PSQSVGT	6
UCRSCGT	10
UNISCGT	12
ENDSVGT	18
ADIUFGT	21
ANESVGT	41
REHSVGT	43
UOGUFGT	44
UNOSCGT	62
DERSVGT	73
UCOSCGT	78
REUSVGT	85
UNESCGT	86
UFIUFGT	91
CMFSVGT	110
MIVSVGT	125
UNRIUFG	127
UGAUFGT	136
GASSVGT	144

Departamento	Ingresos
HEPSCGT	181
CTOSVGT	220
CCASVGT	228
EMESVGT	253
CPLSVGT	277
NFRSVGT	340
ORLSVGT	350
CPESVGT	401
HADUFGT	511
NMLSVGT	558
MIRSVGT	853
NRCSVGT	970
OFTSVGT	993
CARSVGT	1102
ACVSVGT	1295
NRLSVGT	1350
UROSVGT	2685
GINSCGT	3364
OBSSCGT	3907
CGDSVGT	4855
COTSVGT	9870
PEESVGT	13914
URGSVGT	15048

Total:	64844
2,50%	1621

Por tanto, vamos a ignorar del estudio los departamentos marcados en rojo, y limitaremos el estudio a los departamentos con un mayor número de impacto y peso: Urología, ginecología, obstetricia, cirugía general, cirugía traumatológica, pediatría y medicina de urgencias.

Vamos a permitirnos tomar en cuenta el departamento de neumología, porque aunque el número de casos de pequeño, y seguramente no aportará ningún peso decisivo al estudio final, intuimos que podría verse influenciado en gran manera por la gripe, y podría aportar mayor calidad a los datos finales.

Seguiremos ahora analizando la evolución del resto de departamentos en las fechas del estudio.

El mejor método para estudiar la correlación de las series de datos de los departamentos, teniendo en cuenta que cada uno tiene una frecuencia distinta, es mediante el **coeficiente de correlación de Pearson**[2].
Este coeficiente es una herramienta estadística que nos permite establecer relaciones entre dos series de datos.

De esta manera, podemos obtener un coeficiente que nos indique de qué manera ambas series poseen una correlación. Dicho coeficiente variará entre -1 y 1, según si la correlación es directa (1), inversa (-1)...

A valores más cercanos a 1 podremos afirmar que ambas series evolucionan en el tiempo de la misma manera, por lo que podremos evaluar de qué manera los distintos departamentos están afectados por la gripe de igual manera que el departamento de medicina general.

De esta manera, obtenemos el coeficiente de cada una de las series comparándola con el departamento de medicina general. Supondremos que existe una relación apreciable para el estudio cuando su relación sea mayor a 0.50

Por ejemplo, el coeficiente de Pearson de “Neumología” relacionado con medicina general es 0.50, con lo cual podemos postular que ambos departamentos podrían poseer una cierta relación. También podemos apreciar que el departamento de pediatría, que en un principio se podría pensar que podría no estar relacionado, sí que tiene una relación directa con la gripe, incluso mayor que la de neumología.

No obstante, otros departamentos como “Cirugía torácica”, o “obstetricia”, claramente no poseen una relación con directa con la gripe, puesto que sus coeficientes son bastante bajos, aunque positivos.

[2] En este caso, vamos a trabajar con la relación lineal entre dos poblaciones, obteniendo estos valores para cada uno de sus elementos. Para ello, una herramienta estadística muy utilizada es la correlación de Pearson.
http://es.wikipedia.org/wiki/Coeficiente_de_correlaci%C3%B3n_lineal

	Urología dep 47	Ginecología dep 16	Obstetricia dep 27	Cirugía G dep 06	Cirugía T dep 08	Pediatría dep 31	Neumología dep 24	M. Urgencias dep 46
01/01/2012	82	113	182	171	357	484	29	557
08/01/2012	90	118	146	187	399	426	21	547
15/01/2012	105	146	146	211	357	464	22	594
22/01/2012	126	124	154	167	415	589	21	534
29/01/2012	108	126	157	187	372	550	22	520
05/02/2012	98	139	152	191	363	569	23	573
12/02/2012	112	135	147	190	326	634	30	697
19/02/2012	116	133	149	212	395	648	30	730
26/02/2012	99	131	130	170	387	528	23	636
04/03/2012	102	132	168	203	399	526	25	574
11/03/2012	112	134	147	185	381	556	25	555
18/03/2012	93	119	151	191	377	492	10	505
25/03/2012	103	97	140	171	346	496	3	515
30/12/2012	62	53	120	131	228	406	23	439
06/01/2013	95	126	163	156	339	488	25	625
13/01/2013	83	125	159	196	380	538	19	552
20/01/2013	104	125	157	178	378	682	16	612
27/01/2013	117	192	115	197	386	664	25	595
03/02/2013	121	109	198	189	393	531	13	581
10/02/2013	107	136	149	188	414	476	16	617
17/02/2013	107	144	113	205	413	502	25	588
24/02/2013	106	119	147	194	370	471	14	547
03/03/2013	105	131	118	202	405	538	13	554
10/03/2013	105	138	132	176	393	467	12	507
17/03/2013	104	136	123	178	392	503	13	535
24/03/2013	85	100	153	133	395	482	6	544
Coeficiente Pearson:	0,455645	0,48638075	0,1392396	0,43272	0,260847	0,58834	0,5001712	1

Existe un departamento en el cual podemos tener algunas dudas, el de ginecología. No es un apartado que, a priori, pudiera parecer relacionado con la gripe, aunque quizás en este hospital en concreto, llevan un control sobre mujeres embarazadas cuando existe una epidemia.
Dado el peso de este departamento, y su proximidad al valor límite (0.48 frente al 0.50 fijado), lo tendremos en cuenta en un análisis posterior.

Realizamos el estudio tanto con frecuencias semanales como con las diarias, pero los resultados son bastante similares, por lo que sólo mostraremos en este documento las frecuencias semanales, mucho más sencillas de analizar y plasmar en la documentación.

Tras este estudio preliminar, podemos observar que el departamento que posee una correlación mayor con “medicina general” es el de “Pediatría”.
Además, queremos analizar con más detalle si los departamentos de “Ginecología” y “Neumonía” pueden estar influenciados o no, ya que poseen valores límite.

Para convertir la intuición de estos primeros resultados en una certeza, vamos a obtener los datos semanales de búsquedas de google, que utilizaremos como marcador para detectar brotes de gripe, y verificaremos el coeficiente de Pearson de google con medicina general, y el coeficiente si sumamos a los datos de medicina general alguno de los departamentos con mayor correlación.

	Google	M. Urgencias	Urgencias + neumología	Urgencias + pediatría	Urgencias + ginecología
01/01/2012	42	557	586	1041	670
08/01/2012	49	547	568	973	665
15/01/2012	53	594	616	1058	740
22/01/2012	57	534	555	1123	658
29/01/2012	103	520	542	1070	646
05/02/2012	205	573	596	1142	712
12/02/2012	377	697	727	1331	832
19/02/2012	361	730	760	1378	863
26/02/2012	169	636	659	1164	767
04/03/2012	62	574	599	1100	706
11/03/2012	53	555	580	1111	689
18/03/2012	29	505	515	997	624
25/03/2012	24	515	518	1011	612
30/12/2012	91	439	462	845	492
06/01/2013	75	625	650	1113	751
13/01/2013	164	552	571	1090	677
20/01/2013	277	612	628	1294	737
27/01/2013	482	595	620	1259	787
03/02/2013	373	581	594	1112	690
10/02/2013	414	617	633	1093	753
17/02/2013	280	588	613	1090	732
24/02/2013	121	547	561	1018	666
03/03/2013	141	554	567	1092	685
10/03/2013	59	507	519	974	645
17/03/2013	64	535	548	1038	671
24/03/2013	46	544	550	1026	644
Coeficiente Pearson:	1	0,63454202	0,6289153	0,68620676	0,65416337

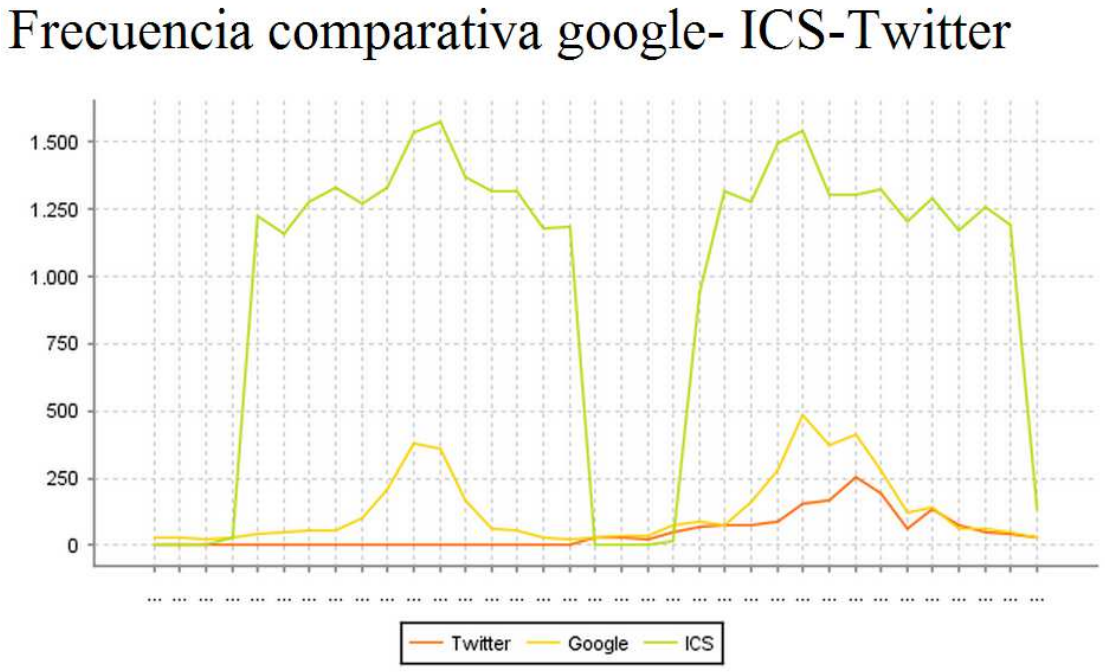
	Google	Urgencias + pediatría + ginecología	Urgencias + neumología + ginecología	Urgencias + pediatría + neumología + ginecología	Urgencias + pediatría + neumología
01/01/2012	42	1154	699	1183	1070
08/01/2012	49	1091	686	1112	994
15/01/2012	53	1204	762	1226	1080
22/01/2012	57	1247	679	1268	1144
29/01/2012	103	1196	668	1218	1092
05/02/2012	205	1281	735	1304	1165
12/02/2012	377	1466	862	1496	1361
19/02/2012	361	1511	893	1541	1408
26/02/2012	169	1295	790	1318	1187
04/03/2012	62	1232	731	1257	1125
11/03/2012	53	1245	714	1270	1136
18/03/2012	29	1116	634	1126	1007
25/03/2012	24	1108	615	1111	1014
30/12/2012	91	898	515	921	868
06/01/2013	75	1239	776	1264	1138
13/01/2013	164	1215	696	1234	1109
20/01/2013	277	1419	753	1435	1310
27/01/2013	482	1451	812	1476	1284
03/02/2013	373	1221	703	1234	1125
10/02/2013	414	1229	769	1245	1109
17/02/2013	280	1234	757	1259	1115
24/02/2013	121	1137	680	1151	1032
03/03/2013	141	1223	698	1236	1105
10/03/2013	59	1112	657	1124	986
17/03/2013	64	1174	684	1187	1051
24/03/2013	46	1126	650	1132	1032
Coeficiente Pearson:	1	**0,68957804**	0,64906579	0,68777042	0,68430677

Tras este segundo análisis, verificamos que añadiendo al departamento de medicina general el de pediatría, la correlación con los datos obtenidos de la red aumenta notablemente. Además, si a estos departamentos sumamos el de ginecología, la correlación crece hasta su máximo (0.689).

Tras estos estudios, podemos concluir que, la mayor correlación de datos entre los proporcionados por el ICS y los obtenidos en google, se obtiene cuando tomamos en consideración los datos de los departamentos de **"Medicina de urgencia", "Ginecología" y "Pediatría" (coeficiente de Pearson = 0.689).** Por tanto, son éstos datos los que tendremos en cuenta para el proyecto.

5.2 Detección de picos de gripe y correlación con otras fuentes.

En este apartado vamos a intentar discernir cuáles son los picos de servicio relacionados con la gripe en la ICS, y qué correlación podemos extrapolar al comparar los estos datos con el resto de fuentes de los que disponemos.
Para este apartado, nos basaremos en el informe de frecuencia comparativa entre los datos de distintas fuentes:



Tras examinar con detalle este informe podemos advertir lo siguiente:

- La media de asistencias en los meses invernales de los que se posee datos, está en torno a las 1250 semanales (Hay que tener en cuenta que estamos contemplando tan sólo los departamentos de medicina de urgencia, pediatría y ginecología)

- Podemos establecer un pico en cada uno de los años, con un número de asistencias que se incrementa en un 20% sobre lo normal, llegando y superando las 1500 asistencias/semanales. Éstos valores se alcanzan en las siguientes semanas:

Fecha	Asistencias
12/02/2012	1466
19/02/2012	1511
20/01/2013	1419
27/01/2013	1451

- Los datos que poseemos de twitter (tan sólo en el año 2013) poseen un desplazamiento en el tiempo respecto a los picos de gripe. ¿Por qué podría ser debido este comportamiento? Podríamos pensar que, cuando una persona posee un caso agudo de gripe que requiera asistencia, lo postea en redes sociales a posteriori, comentando que está o ha estado enfermo (hemos de tener en cuenta que el período de convalecencia de esta enfermedad suele ser entre 10-14 días). Ello podría explicar el desplazamiento de 1-2 semanas que existe entre los mayores picos de frecuencia en el hospital y el incremento de tuits.
  Nos haría falta un mayor muestreo de datos para confirmar estas suposiciones, pero, con la información que poseemos, **debemos descartar twitter como fuente de información para predecir brotes de gripe.**

- La otra fuente de información de la que disponemos es Google (datos provenientes de google flu).
  En este caso sí que parece que existe una relación directa con los datos del ICS, y que se produce en el mismo intervalo de tiempo.
  ¿A podría deberse esta correlación instantánea?
  Quizás a que tradicionalmente se ha utilizado google como herramienta de consulta ante algún problema, al contrario que las redes sociales, donde usualmente se suelen publicar hechos ya acaecidos. Por ello, los usuarios deben estar utilizando en primer lugar el buscador para obtener información puntual sobre su estado de salud, y de esta manera se obtienen estos datos sobre consultas de manera inmediata.
  Lo importante en este asunto es que sí que podemos establecer una relación directa e inmediata entre ambas fuentes, por lo que podemos utilizar con relativa fiabilidad los datos publicados por el buscador para predecir con razonables garantías futuros brotes de gripe.

Respecto a los indicadores que nos pueden indicar cuándo estamos ante un brote de gripe, podemos situar con una razonable seguridad el límite en las 250 consultas. No obstante, esa información la tenemos **a posteriori**, por lo que hay que buscar otras formas de intentar adivinar este brote con los datos que disponemos.

Es muy complicado extraer este tipo de conclusiones con tan pocos datos.
Hay que fijarse que, aunque tenemos información de dos años, realmente tan sólo podemos constatar la existencia de 2 brotes de gripe.

	Consultas Google	Asistencias ICS
01/01/2012	42	1154
08/01/2012	49	1091
15/01/2012	53	1204
22/01/2012	57	1247
29/01/2012	103	1196
05/02/2012	205	1281
12/02/2012	377	1466
19/02/2012	361	1511
26/02/2012	169	1295
04/03/2012	62	1232
11/03/2012	53	1245
18/03/2012	29	1116
25/03/2012	24	1108
30/12/2012	91	898
06/01/2013	75	1239
13/01/2013	164	1215
20/01/2013	277	1419
27/01/2013	482	1451
03/02/2013	373	1221
10/02/2013	414	1229
17/02/2013	280	1234
24/02/2013	121	1137
03/03/2013	141	1223
10/03/2013	59	1112
17/03/2013	64	1174
24/03/2013	46	1126

La información que tenemos de estos brotes es la siguiente:
-Siempre (2 veces de 2), han durado 2 semanas.
-El período anterior a un brote de gripe, el número de consultas en google se casi duplicó, superándose además la media aritmética de consultas (110)
En la tabla anterior, podemos identificar la semana del 05/02/2012 como la previa al brote, mientras que la del 12/02/2012 y 19/02/2012 corresponden a la del brote.
Nuestra máxima prioridad será intentar predecir el brote antes de que se produzca, por lo que intentaremos conocer de antemano qué semana va a ser la previa a dicho brote.

Por tanto, los criterios que utilizaremos para predecir futuros brotes de gripe serán los ya indicados: que se supere la media aritmética de consultas y que exista un importante incremento en el número de consultas. Ello, debido a la correlación existente entre los datos de google y los del ICS, nos indicará que existe una razonable seguridad de que se produzca un importante repunte en los casos de gripe en el período posterior.
¿Y qué probabilidad tenemos de que se produzca este brote? En el mejor de los casos, la probabilidad de predecirlo coincidirá con el coeficiente de correlación entre ambas series **(coeficiente de Pearson), es decir un 68,9%**

Además, una vez producido un brote, asignaremos una probabilidad del **100%** a que también exista en el período siguiente.
Como en todo estudio estadístico, estos datos irán refinándose y aumentando su fiabilidad conforme se disponga de más valores. Por ello es importante ir almacenando los valores de años posteriores, para que la validez del sistema se incremente con el tiempo.

6 Implementación.

En esta fase, realizaremos la implementación de las fases definidas en el proceso de diseño: carga, transformación, creación de informes y cuadro de mandos.
Además, se anexarán también en el apartado siguiente las pruebas realizadas, por lo que estos dos puntos son los más técnicos de la monografía. No obstante, se puede afirmar que se recogen todas las consideraciones previas, y que en estas fases tan sólo estamos implementando el diseño preliminar en código.

6.1 Fase de carga de datos.

En esta fase, obtenemos los datos de distintas fuentes y los incorporamos a la base de datos ya creada.
Dado la variabilidad de las fuentes de datos, y al objetivo de que el proyecto sea lo más modular y escalable posible, se ha optado por realizar procedimientos semi-automatizados para cargar los datos.
Ello nos proporciona las siguientes ventajas:

-Relativa independencia respecto al formato de los datos de entrada.
-Indiferencia respecto al sistema operativo utilizado, ya que la automatización se realiza con ANSI-SQL.
-Reutilización y reusabilidad de los procedimientos.
-Fácil adaptabilidad de los procedimientos ante cambios.

Por todo ello, se ha preferido la realización de procedimientos bien documentados a automatismos que aportarían poco al trabajo final y dificultarían cualquier modificación aunque fuera pequeña.

Según la fase de análisis, obtendremos los datos necesarios para el proyecto de diversas fuentes:

-Datos provenientes de ICS. Estos datos son los más importantes, puesto que nos van a servir para compararlos con el resto de fuentes e intentar predecir brotes de gripe.
Dichos datos se nos facilitan de partida, en formato excel.
Para cargar estos datos en nuestra BD, crearemos una sentencia ANSI SQL que nos permita cargarlos directamente.
Por tanto el procedimiento para cargar estos datos es el siguiente:

1.-A partir de la hoja excel “BI_UOC_Datos_Hospitalizacionv1.xls”, grabar en formato “csv” y sin cabeceras las columnas “Data Inici Hosp Trunc”, “Servei Hospitalitzacio”,” Diagnostic P (desc)”.

2.- Ejecutar la siguiente Query en la base de datos MySql:

-- ----------

```
-- Carga  TC_ORI_HOSPITAL
-- ----------------------------------------------------
DELETE FROM mydb.tc_ori_hospital;
LOAD DATA LOCAL INFILE '../../../../tmp/hospital.csv' INTO TABLE mydb.tc_ori_hospital
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
(@col1,@col2,@col3)
set
Fecha_Ingreso = str_to_date(@col1, '%d/%m/%Y'),
Servicio = @col2,
Descripcion_diagnostico = @col3;
```

Tras estos pasos, habremos cargado la tabla TC_ORI_HOSPITAL.
Reseñar que la primera línea borra la tabla, por lo que si se trata de sucesivas incorporaciones de datos, habrá que eliminar la primera instrucción y realizar tan sólo la segunda.

3.-A partir de la hoja excel “CasBI_UOC_v1Urgencias.xlsx”, grabar en formato “csv” y sin cabeceras las columnas “Data Entrada”, “Servei Entrada”,” Servei Alta”.

4.- Ejecutar la siguiente Query en la base de datos MySql:

```
-- ----------------------------------------------------
-- Carga  TC_ORI_URGENCIAS
-- ----------------------------------------------------
DELETE FROM mydb.tc_ori_urgencias;
LOAD DATA LOCAL INFILE '../../../../tmp/urgencias.csv' INTO TABLE mydb.tc_ori_urgencias
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
(@col1,@col2,@col3)
set
Fecha_Entrada = str_to_date(@col1, '%d/%m/%Y'),
Servicio_Entrada = @col2,
Servicio_Alta = @col3;
```

Tras estos pasos, habremos cargado la tabla TC_ORI_URGENCIAS.
Reseñar que la primera línea borra la tabla, por lo que si se trata de sucesivas incorporaciones de datos, habrá que eliminar la primera instrucción y realizar tan sólo la segunda.

5.- Cargamos los siguientes datos correspondientes a los distintos departamentos.
Estos datos se podrían obtener desde las tablas excel anteriores, pero consideramos que deberíamos incorporarlo como datos fijos, ya que la existencia de un departamento no ha de supeditarse a que aparezca en los datos que disponemos puntualmente. Ello podría provocar que posteriores incorporaciones de datos eliminaran información de departamentos ya

existentes, o bien no asignaran las mismas claves a los mismos departamentos, por lo que es más correcto informar dichos datos como valores constantes

```
-- ----------------------------------------------------
-- Carga  TW_WRK_SERVICIO
-- ----------------------------------------------------

DELETE FROM mydb.tw_wrk_servicio;
INSERT INTO mydb.tw_wrk_servicio VALUES (01,"ACVSVGT","ANGIOLOGIA I CIRURGIA VASCULAR");
INSERT INTO mydb.tw_wrk_servicio VALUES (02,"ADIUFGT","UNITAT ADDICCIONS - DESINTOXIC");
INSERT INTO mydb.tw_wrk_servicio VALUES (03,"ANCSVGT","SIN DESCRIPCION DISPONIBLE    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (04,"ANESVGT","ANESTESIOLOGIA I REANIMACIO   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (05,"APASVGT","SIN DESCRIPCION DISPONIBLE    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (06,"CARSVGT","CARDIOLOGIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (07,"CCASVGT","CIRURGIA CARDIACA             ");
INSERT INTO mydb.tw_wrk_servicio VALUES (08,"CGDSVGT","CIRURGIA GENERAL I DIGESTIVA  ");
INSERT INTO mydb.tw_wrk_servicio VALUES (09,"CMFSVGT","CIRURGIA ORAL I MAXIL.LOFACIAL");
INSERT INTO mydb.tw_wrk_servicio VALUES (10,"COTSVGT","CIRURGIA ORT I TRAUMATOLOGIA  ");
INSERT INTO mydb.tw_wrk_servicio VALUES (11,"CPESVGT","CIRURGIA PEDIATRICA           ");
INSERT INTO mydb.tw_wrk_servicio VALUES (12,"CPLSVGT","CIRURGIA PLASTICA I REPARADORA");
INSERT INTO mydb.tw_wrk_servicio VALUES (13,"CTOSVGT","CIRURGIA TORACICA             ");
INSERT INTO mydb.tw_wrk_servicio VALUES (14,"DERSVGT","DERMATOLOGIA MEDICO-QUIR I VEN");
INSERT INTO mydb.tw_wrk_servicio VALUES (15,"EMESVGT","TRASLLAT                      ");
INSERT INTO mydb.tw_wrk_servicio VALUES (16,"ENDSVGT","ENDOCRINOLOGIA I NUTRICIO     ");
INSERT INTO mydb.tw_wrk_servicio VALUES (17,"GASSVGT","GASTROENTEROLOGIA- AP DIGESTIU");
INSERT INTO mydb.tw_wrk_servicio VALUES (18,"GINSCGT","GINECOLOGIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (19,"HADUFGT","UNITAT D'HOSPITALITZACIO A DOM");
INSERT INTO mydb.tw_wrk_servicio VALUES (20,"HEMSVGT","HEMATOLOGIA CLINICA           ");
INSERT INTO mydb.tw_wrk_servicio VALUES (21,"HEPSCGT","HEPATOLOGIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (22,"HIVUFGT","UNITAT HIV                    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (23,"IMESVGT","SIN DESCRIPCION DISPONIBLE    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (24,"MIRSVGT","MEDICINA INTERNA              ");
INSERT INTO mydb.tw_wrk_servicio VALUES (25,"MIVSVGT","MEDICINA INTENSIVA            ");
INSERT INTO mydb.tw_wrk_servicio VALUES (26,"NFRSVGT","NEFROLOGIA                    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (27,"NMLSVGT","PNEUMOLOGIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (28,"NRCSVGT","NEUROCIRURGIA                 ");
INSERT INTO mydb.tw_wrk_servicio VALUES (29,"NRLSVGT","NEUROLOGIA                    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (30,"OBSSCGT","OBSTETRICIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (31,"OFTSVGT","OFTALMOLOGIA                  ");
INSERT INTO mydb.tw_wrk_servicio VALUES (32,"ONCSVGT","ONCOLOGIA MEDICA              ");
INSERT INTO mydb.tw_wrk_servicio VALUES (33,"ORLSVGT","OTORRINOLARINGOLOGIA          ");
INSERT INTO mydb.tw_wrk_servicio VALUES (34,"PEESVGT","PEDIATRIA                     ");
INSERT INTO mydb.tw_wrk_servicio VALUES (35,"PSQSVGT","PSIQUIATRIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (36,"REHSVGT","MEDICINA FISICA REHABILITACIO ");
INSERT INTO mydb.tw_wrk_servicio VALUES (37,"REUSVGT","REUMATOLOGIA                  ");
INSERT INTO mydb.tw_wrk_servicio VALUES (38,"UCOSCGT","UNITAT CORONARIA              ");
INSERT INTO mydb.tw_wrk_servicio VALUES (39,"UCRSCGT","UNITAT DE CRITICS             ");
INSERT INTO mydb.tw_wrk_servicio VALUES (40,"UFIUFGT","UFISS / GERIATRIA             ");
INSERT INTO mydb.tw_wrk_servicio VALUES (41,"UGAUFGT","UNITAT DE GERIATRIA AGUT      ");
INSERT INTO mydb.tw_wrk_servicio VALUES (42,"UNESCGT","NEONATOLOGIA-CURES SEMICRITIC ");
INSERT INTO mydb.tw_wrk_servicio VALUES (43,"UNISCGT","NEONATOLOGIA-CURES INTENSIVES ");
INSERT INTO mydb.tw_wrk_servicio VALUES (44,"UNOSCGT","NEONATOLOGIA-OBSERVACIO       ");
INSERT INTO mydb.tw_wrk_servicio VALUES (45,"UNRIUFG","UNITAT NEURORADIOLOGIA INTERVE");
INSERT INTO mydb.tw_wrk_servicio VALUES (46,"UOGUFGT","UNITAT DE ORTOGERIATRIA       ");
INSERT INTO mydb.tw_wrk_servicio VALUES (47,"UOMUFGT","UNITAT D'OBESITAT MORBIDA     ");
INSERT INTO mydb.tw_wrk_servicio VALUES (48,"UPMUFGT","UNITAT PATOLOGIA MAMA         ");
INSERT INTO mydb.tw_wrk_servicio VALUES (49,"URGSVGT","MEDICINA D'URGENCIES          ");
INSERT INTO mydb.tw_wrk_servicio VALUES (50,"UROSVGT","UROLOGIA                      ");
```

-Datos provenientes de google.

El procedimiento es muy similar al anterior. Descargaremos la información, puesto que es pública y se actualiza frecuentemente, y realizaremos los siguientes procedimientos para cargarlos en nuestra base de datos:

1.- Descargar los datos de google flu: http://www.google.org/flutrends/es/data.txt:
Importar los datos a un fichero csv, eliminando las cabeceras, y dejando los datos de Cataluña.

2.- Ejecutar la siguiente Query en la base de datos MySql:

```
-- ------------------------------------------------------
-- Carga  TC_ORI_GOOGLE
-- ------------------------------------------------------
DELETE FROM mydb.tc_ori_google;
LOAD DATA LOCAL INFILE '../../../../tmp/google_flu_cat.csv' INTO TABLE mydb.tc_ori_google
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
(@col1,@col2)
set
Fecha_Intervalo = str_to_date(@col1, '%d/%m/%Y'),
Frecuencia = @col2;
```

Tras estos pasos, habremos cargado la tabla TC_ORI_GOOGLE.
Al contrario que en el resto de casos, los datos ofrecidos por google por ahora incluyen todos los datos obtenidos desde que se comenzó la estadística, por lo que para actualizar nuestros tablas habrá que seguir borrando la tabla destino antes de volver a cargar los datos.

-Datos provenientes de twitter.
En el análisis no se ha podido encontrar una solución viable para obtener tuits con una antigüedad superior a 7 días.
No obstante, se prepara la infraestructura de la aplicación para poder incorporar los datos que sea posible, bien por encontrar algún servicio especializado que pueda suministrarnos estos datos, bien nosotros mismos mediante la descarga diaria de tuits hasta tener una cantidad estadísticamente significativa.
Para este apartado, se ha conseguido información de años anteriores en un fichero csv, con lo cual vamos a preparar un procedimiento similar al seguido con el resto de datos.

1.-A partir del fichero csv “tweets_gripe_2013.csv”, grabar sin cabeceras las columnas “Contenido”, “Keyword”,” “Fecha de tweet”.

2.- Ejecutar la siguiente Query en la base de datos MySql:

```
-- ----------------------------------------------------
-- Carga  TC_ORI_TWITTER
-- ----------------------------------------------------
DELETE FROM mydb.tc_ori_twitter;
LOAD DATA LOCAL INFILE '../../../../tmp/tweets.csv' INTO TABLE mydb.tc_ori_twitter
FIELDS TERMINATED BY ';'
LINES TERMINATED BY '\n'
(@col1,@col2,@col3,@col4)
set
Tuit = @col1,
Busqueda = @col2,
Fecha = str_to_date(@col3, '%d/%m/%Y'),
Idioma = @col4;
```

Tras estos pasos, habremos cargado la tabla TC_ORI_TWITTER.
Como en otros casos, observamos que la primera línea borra la tabla, por lo que si se trata de sucesivas incorporaciones de datos, habrá que eliminar la primera instrucción y realizar tan sólo la segunda.

6.2 Fase de transformación de datos

En esta fase, vamos a convertir los datos crudos en información elaborada que nos servirá de base para informes y cuadros de mando.
Al principio de esta fase, tenemos la información inicial sin modificaciones. Se ha excluido parte de la información que se ha considerado no relevante, pero no se ha tratado dato alguno.
Al final de la fase de transformación de datos, obtendremos la información relevante en el formato adecuado para su proceso.

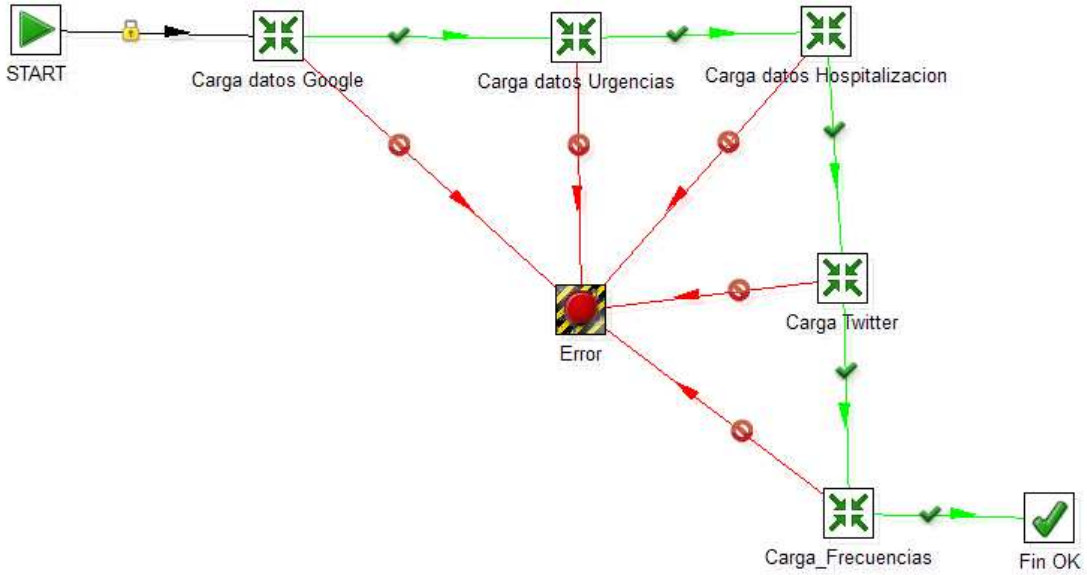
Para realizar esta transformación se va a utilizar la herramienta ETL de Pentaho, denominada “Spoon”.

La nomenclatura de la base de datos utilizada es la siguiente:
-Las tablas que comienzan por “TC_” significan “Tabla de Carga”. En ellas se guardan los datos crudos, recién cargados en la base de datos.
Será en estas tablas donde podremos revisar los datos cargados, estando siempre disponibles, y dispuestos para ser consultados para revisar los criterios de selección utilizados. Se guarda el máximo de información útil en estas tablas, pero ningún dato que pueda ser sensible de ser protegido por LOPD.
Ninguna de estas tablas van a ser utilizadas en pasos posteriores, por lo que al terminar el proceso de transformación, y una vez validados los datos, pueden ser borradas sin problemas.

-Las tablas que comienzan por “TW_WRK_” son el segundo nivel cargado. Se trata de tablas de trabajo del datawarehouse, con información que ha comenzado a ser tratada pero no está en su formato final. Al igual que en el caso anterior, se guarda información adicional en estas tablas (como el texto del tuit, o el diagnóstico médico) para poder revisar los criterios utilizados.
Existen informes asociados a estas tablas, para poder realizar tareas de control, pero en teoría, y una vez dados por buenos estos datos, estas tablas podrían borrarse en el caso de que el espacio fuera un problema.

-Por último, las tablas que comienzan por “TW_WAR_” ya poseen información tal y como va a ser utilizada en el cuadro de mandos, y es la información básica sobre la que va a sustentarse la aplicación.

-Esquema general de la transformación de datos:

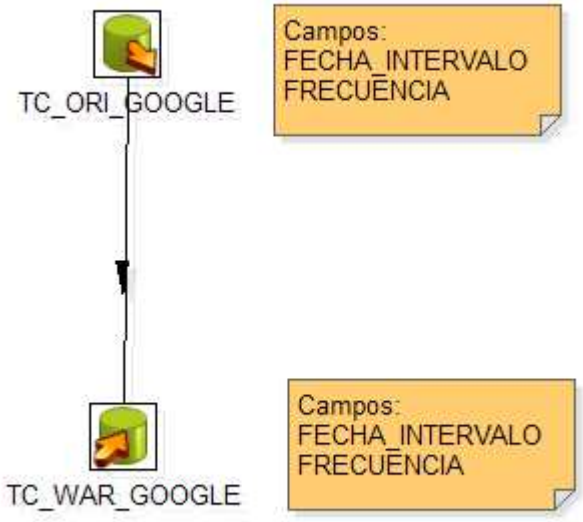


Este esquema es el flujo que seguirá la transformación de datos de la aplicación.
Como se puede apreciar, realizaremos por este orden, las siguientes fases:
-Carga de datos de google, desde la tabla TC_ORI_GOOGLE a TW_WAR_GOOGLE
-Carga de datos de urgencias, desde la tabla TC_ORI_URGENCIAS a TW_WRK_ICS
-Carga de datos de hospitalización, desde la tabla TC_ORI_HOSPITAL a las tablas TW_WRK_ICS y TW_WRK_DIAGNOSTICO
-Carga de datos de twitter, desde la tabla TC_ORI_TWITTER a TW_WRK_TWITTER y TW_WRK_TUITS.
-Adaptación de las frecuencias diarias a semanales de las tablas TW_WRK_ICS y TW_WRK_TUITTER a las tablas TW_WAR_ICS y TW_WAR_TUITTER

Podemos apreciar como cualquier error en uno de estos pasos provocará que el proceso completo se detenga.
Seguidamente vamos a ver con más detalle cómo se realiza cada uno de estos pasos.

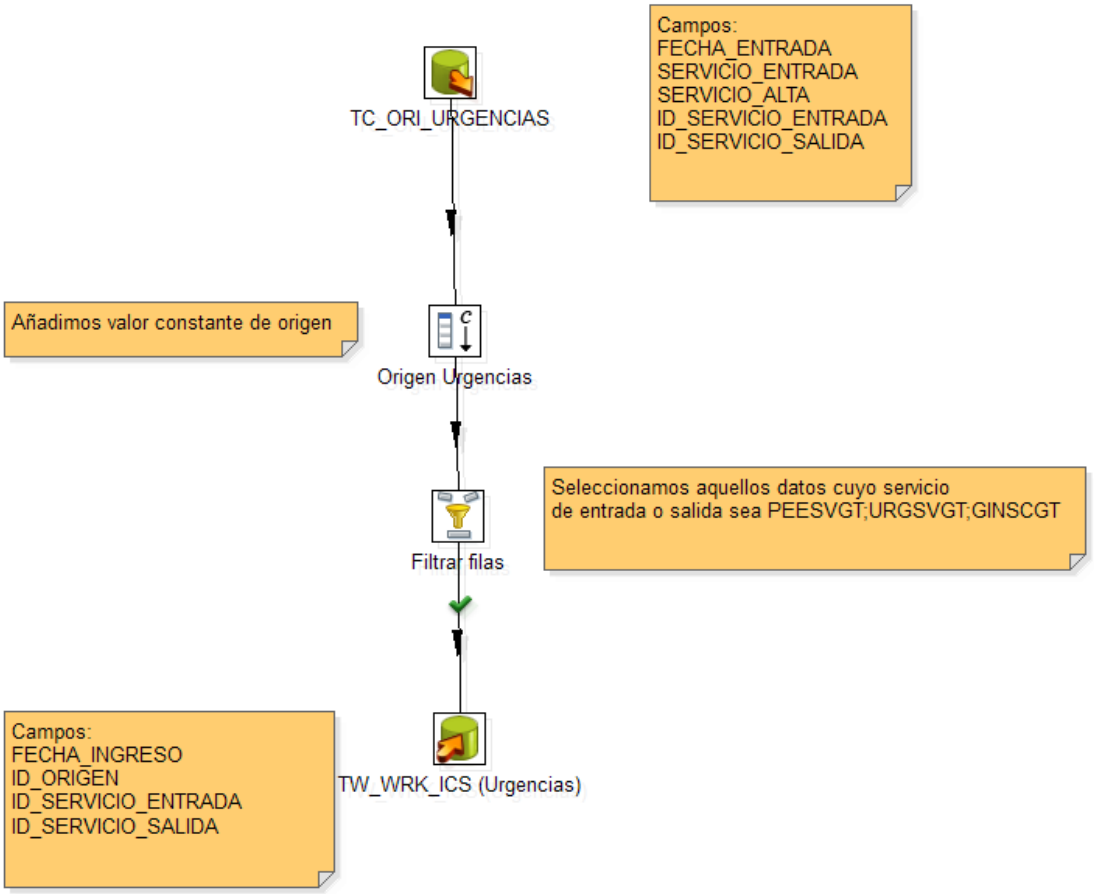
-Carga de datos de google.
En este paso vamos a realizar la carga de datos de google, desde la tabla TC_ORI_GOOGLE a TW_WAR_GOOGLE

Es, de lejos, la transformación más sencilla de todo el proceso, ya que la estructura de la tabla TC_ORI_GOOGLE es análoga a la de TW_WAR_GOOGLE.
Por tanto, este proceso se limitará a pasar los datos de una tabla a otra, respetando la estructura de campos.



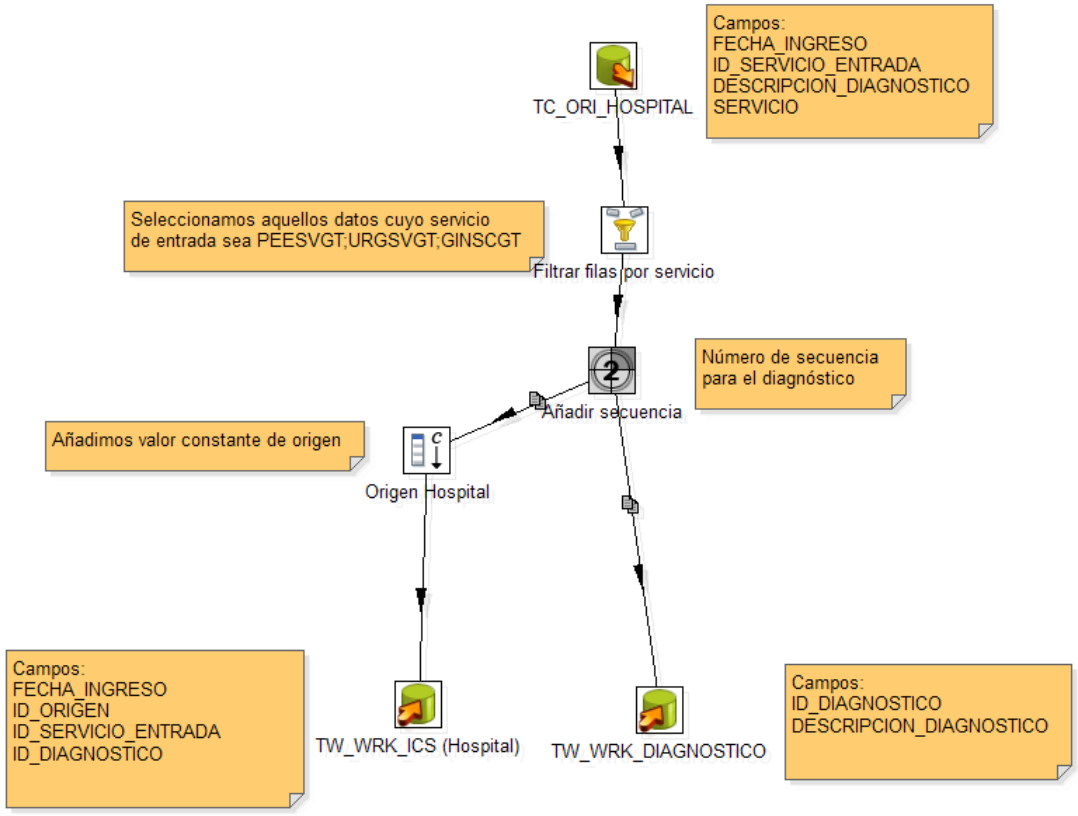
-Carga de datos de urgencias.
En este paso vamos a realizar la carga de datos de urgencias, desde la tabla TC_ORI_URGENCIAS a TW_WRK_ICS

El procedimiento será el siguiente:
-Seleccionamos la fecha, los servicios de entrada y salida y sus correspondientes identificadores numéricos.
-Añadimos un campo constante para identificar el origen del dato.
-Realizamos un filtrado para seleccionar aquellos servicios que tienen más relevancia con la gripe (ver el capítulo de **Validez y precisión de los datos**) : Pediatría, ginecología y medicina de urgencias.
-Los datos resultantes se graban en la tabla TW_WRK_ICS.


Campos:
FECHA_ENTRADA
SERVICIO_ENTRADA
SERVICIO_ALTA
ID_SERVICIO_ENTRADA
ID_SERVICIO_SALIDA
TC_ORI_URGENCIAS
Añadimos valor constante de origen
Origen Urgencias
Seleccionamos aquellos datos cuyo servicio
de entrada o salida sea PEESVGT;URGSVGT;GINSCGT
Filtrar filas
Campos:
FECHA_INGRESO
ID_ORIGEN
ID_SERVICIO_ENTRADA
ID_SERVICIO_SALIDA
TW_WRK_ICS (Urgencias)

-Carga de datos de hospitalizaciones.

En este paso vamos a realizar la carga de datos de hospitalizaciones, desde la tabla TC_ORI_HOSPITAL a las tablas TW_WRK_ICS y TW_WRK_DIAGNOSTICO.



El procedimiento será el siguiente:

-Seleccionamos la fecha, los servicios de entrada, sus correspondiente identificador numérico y la descripción del diagnóstico.

-Realizamos un filtrado para seleccionar aquellos servicios que creemos tienen más relevancia con la gripe : Pediatría, ginecología y medicina de urgencias. (ver el capítulo de **Validez y precisión de los datos**)

-Incluimos un número secuencial para que actúe de campo clave del diagnóstico.

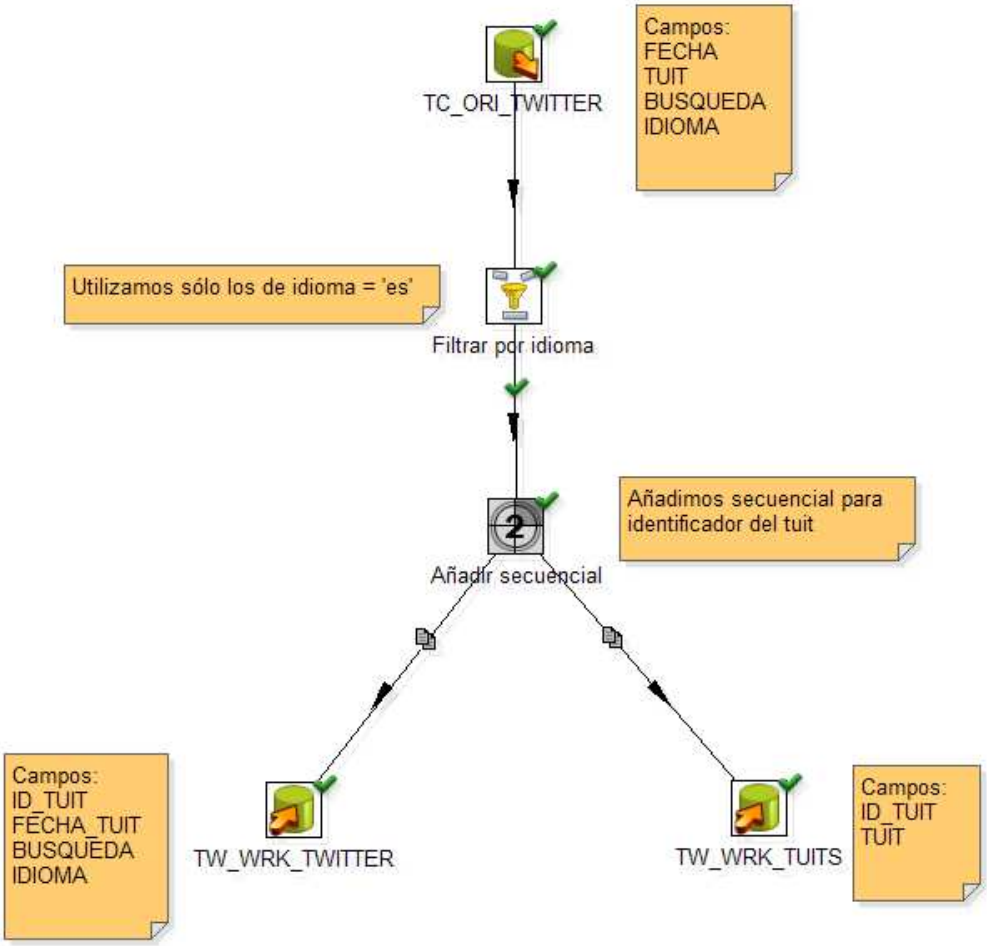
-Añadimos un campo constante para identificar el origen del dato.

-Grabamos en la tabla TW_WRK_ICS los campos Fecha_ingreso, Id_origen, Id_Servicio_entrada e Id_Diagnostico. Hay que reseñar que estos datos se añaden a los ya cargados anteriormente con los datos de urgencias.

-Grabamos en la tabla TW_WRK_DIAGNOSTICO los campos Id_diagnostico, Descripción_Diagnóstico.

-Carga de datos de Twitter.

En este paso vamos a realizar la carga de datos de twitter, desde la tabla TC_ORI_TWITTER a las tablas TW_WRK_TWITTER y TW_WRK_TUITS.



El procedimiento será el siguiente:

-Seleccionamos la fecha, el tuit, los criterios de búsqueda y el idioma.

-Realizamos un filtrado para seleccionar aquellas palabras clave que creemos tienen más relevancia con la gripe : lenguaje español("es").

-Incluimos un número secuencial para que actúe de campo clave del tuit.

-Grabamos en la tabla TW_WRK_ TWITTER los campos Id_tuit, Fecha_tuit, Busqueda e Idioma.

-Grabamos en la tabla TW_WRK_TUITS los campos Id_tuit, y Tuit.

-Reformateamos los datos para generar una adecuada coherencia.

Una vez cargados todos los datos, observamos que existe una discrepancia entre los criterios seguidos en las distintas fuentes. Este problema es muy común a la hora de crear un datawarehouse, donde nos podemos encontrar distintos criterios en la recogida de datos.

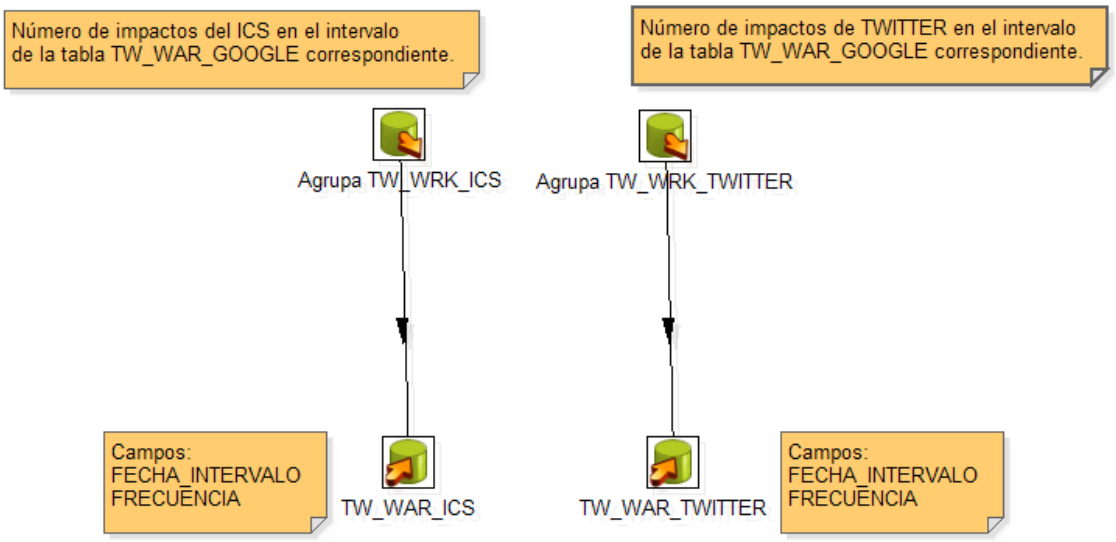
Revisando en detalle, observamos en que los datos obtenidos de google se han seguido el criterio de guardar el número de impactos semanales relacionados con la gripe.

En cambio, en los datos que tenemos de twitter, y del ICS, los datos se han medido de manera diaria.

Es por ello que necesitamos obtener un criterio común para poder tratar y procesar dichos datos.

Como va a ser imposible, o al menos muy poco exacto, convertir los datos de google a una frecuencia diaria (dividiendo el número de impacto por día del intervalo, por ejemplo), pensamos que es mucho más coherente y adecuado obtener la frecuencia semanal de los datos provenientes de Twitter y el ICS.

De esta manera, cargamos las tablas TW_WAR_TWITTER y TW_WAR_ICS obteniendo el número de impactos en cada uno de los intervalos que poseemos de google, y usando la fecha de esta última tabla como referencia.



6.3 Creación de informes.

Se contempla la creación de informes con dos posibles usos:
1.- Revisar la información que se está teniendo en consideración.
Es fácil que los datos puedan quedar falseados si la selección no es correcta, por lo que se plantea realizar una serie de informes que nos aporten información sobre dichos datos.
2.- Obtener una información histórica que demuestre la validez del proyecto.

Por tanto, se consideran útiles los siguientes informes:

Diagnósticos utilizados por fecha y servicio.
Con este informe se muestra un listado con los servicios y los diagnósticos de cada dato concerniente a la hospitalización de pacientes.
Los datos de este informe son los que realmente estamos utilizando en el proyecto, por lo que la principal utilidad de este informe reside en la posibilidad de advertir si estamos obteniendo los datos apropiados para el estudio.
Por ejemplo, la imagen adjuntada podría generar dudas sobre la idoneidad de utilizar los datos del servicio de neumología, ya que parece que también se incluyen datos sobre fracturas que afectan a la cavidad torácica. No obstante, si seguimos observando los datos, podremos concluir que la selección es válida, ya que la mayoría de casos se corresponden con problemas respiratorios relacionados con infecciones víricas.

mayo 16, 2014 @ 05:45

Diagnósticos utilizados por fecha y servicio.

Fecha Ingreso : Oct 15, 2011

Descripción del servicio	Descrición del diagnóstico
PNEUMOLOGIA	Actinomicosi pulmonar. Actinomicosi toracica

Fecha Ingreso : Dec 1, 2011

Descripción del servicio	Descrición del diagnóstico
PNEUMOLOGIA	Fractura patologica de vertebra. Esclafament de vertebra NOS
PNEUMOLOGIA	Neoplasia maligna secundaria retroperitoneu i peritoneu

Fecha Ingreso : Dec 12, 2011

Descripción del servicio	Descrición del diagnóstico
PNEUMOLOGIA	Bronquitis cronica obstructiva amb bronquitis aguda

Frecuencia de atenciones por servicio y fecha.
En este informe en concreto estamos seleccionando todos los datos que poseemos, y los mostramos por intervalo semanal y tipo de servicio.
A diferencia del informe anterior, en el que tan sólo mostrábamos los datos seleccionados para el estudio, en este informe poseemos los datos de todas las especialidades.
Por tanto, la principal utilidad de este informe es verificar, de manera rápida, si algún servicio debería ser incluido en el informe.
Si advertimos que algún servicio incrementa notablemente su frecuencia en los meses de invierno, sería interesante verificar si esta frecuencia está relacionada con la gripe.

Como se puede ver, este informe, junto con el anterior, están pensados para seleccionar con garantías aquellos servicios que pueden ser de mayor utilidad para el proyecto.

mayo 16, 2014 @ 06:07

Frecuencia de atenciones por servicio y fecha

Fecha Intervalo : 09-oct-2011

Descripción del Servicio	Frecuencia
PNEUMOLOGIA	1

Fecha Intervalo : 27-nov-2011

Descripción del Servicio	Frecuencia
PNEUMOLOGIA	2

Fecha Intervalo : 11-dic-2011

Descripción del Servicio	Frecuencia
PNEUMOLOGIA	4

Fecha Intervalo : 18-dic-2011

Descripción del Servicio	Frecuencia
PNEUMOLOGIA	7

Fecha Intervalo : 25-dic-2011

Descripción del Servicio	Frecuencia
PNEUMOLOGIA	18
MEDICINA D'URGENCIES	20

Fecha Intervalo : 01-ene-2012

Descripción del Servicio	Frecuencia
ANGIOLOGIA I CIRURGIA VASCULAR	1
CIRURGIA GENERAL I DIGESTIVA	7
CIRURGIA ORT I TRAUMATOLOGIA	16

Listado de tuits por búsqueda e idioma.

Al incorporar información de twitter como fuente de datos de entrada, es necesario realizar algún tipo de verificación sobre dicha información.

En este informe podremos comprobar si los filtros de idioma y palabras de búsqueda son los adecuados, o hay que introducir cambios.

De esta manera, al recuperar información con diferentes palabras de búsqueda podemos revisar los datos incorporados para verificar si hemos acertado con dichos parámetros.

junio 05, 2014 @ 05:38

Listado de tuits por búsqueda e idioma

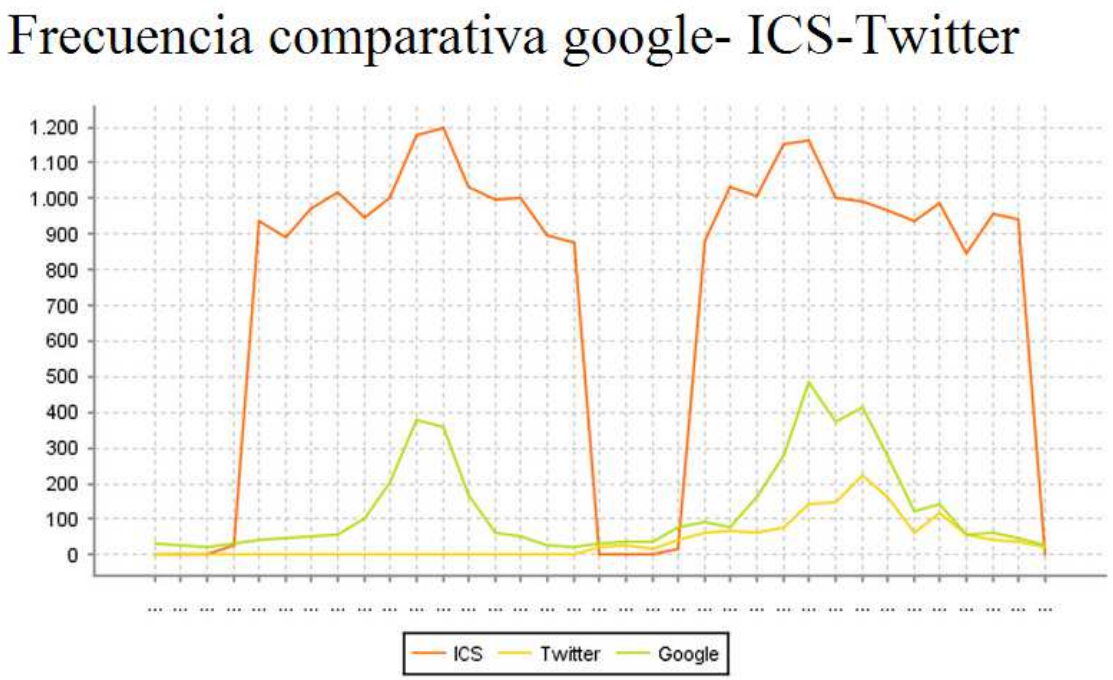
Palabras de búsqueda: gripe

Idioma tuit: ''
''

fecha	tuit
Jan 25, 2013	Por lo visto todos luego estamos con gripe y con la garganta hecho mierda :O
Jan 26, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Jan 30, 2013	Gripe o qu
Jan 30, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Jan 31, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Feb 1, 2013	El ventilador en 1 me da calor, pero en 2 ya me da fr?o. Eso quiere decir q sigo con fiebre o q soy hinchapelotas, nom?s? #duda #gripe
Feb 1, 2013	No s? si es peor esta gripe o verla a usted JAJAJA.
Feb 1, 2013	Ay gripe HDP me jodiste mi fin de semana fiesta!!!:O
Feb 3, 2013	el t? que estoy tomando es un asco, pero quiero que se me pase la gripe o lo que mierda tenga
Feb 5, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Feb 19, 2013	Que te mejore la gripe un d?a y al siguiente estar peor que hace dos &lt
Feb 19, 2013	@JoseAgd14
Feb 21, 2013	@leegchannie bien con mucho color de cabeza y creo que me va a dar gripe... todav?a no empese con exames pero la otra semana seguro.ji.
Apr 2, 2013	Se acerca el finde y yo con gripe...?

Frecuencia comparativa Google-ICS-Twitter.
Este informe es especialmente importante para los desarrolladores, puesto que nos aporta importante información sobre las relaciones de las apariciones de la gripe en cada uno de los medios con la frecuencia de utilización de los servicios sanitarios del ICS.
Se compone de una simple gráfica donde se muestran las frecuencias semanales de cada una de las fuentes.
Esta gráfica se construye con los siguientes datos:
-Datos provenientes de google (google flu)
-Datos recogidos de twitter (búsqueda: gripe, lenguaje: es)
-Datos del ICS de los departamentos seleccionados en el apartado "Validez y precisión de los datos".

Hay que advertir que este informe es válido tan sólo si se han seleccionado los departamentos correctos para hacer la comparación. De otra manera, la información aportada por esta gráfica sería confusa y no nos permitiría observar ninguna correlación.



Este informe será la base del apartado "Detección de picos de gripe y correlación con otras fuentes."

6.4 Creación de un cuadro de mando.

La creación del cuadro de mando seguirá el diseño comentado en el apartado Diseño del cuadro de mandos.

Planteamos la generación del cuadro de mandos con los siguientes requerimientos técnicos:

-Consideramos interesante aportar una visión de los datos en los que nos basamos para inferir las predicciones. Por ello, es deseable incorporar en el cuadro una gráfica de los datos obtenidos de Google y del ICS, para que se pueda apreciar de manera clara la relación entre ellos, y verificar la validez del modelo.
Hay que tener en cuenta que, según el capítulo de **Validez y precisión de los datos,** hemos descartado los datos de twitter, por lo que no deben aparecer en el cuadro de mandos.

-Es imprescindible mostrar los datos más recientes para su estudio, por lo que se debe permitir al usuario revisar de manera visual y rápida las cifras más recientes de las que se disponen.

-En base a los datos anteriores, hay que mostrar de manera visual y de forma destacada, la predicción sobre la posibilidad de que exista un brote de gripe en la semana posterior a la revisada.
Los indicadores (KPI) utilizados para inferir esta probabilidad, se han visto en el apartado Detección de picos de gripe y correlación con otras fuentes., y se postulan como los siguientes:

KPI	Probabilidad
Frecuencia mensual de consultas inferior a la media	Baja
Frecuencia mensual de consultas superior a la media pero inferior a la frecuencia de la semana anterior *1.75	Media
Frecuencia mensual de consultas superior a la media y a 1.75 veces la frecuencia de la semana anterior	Alta. Es muy probable que comience un brote de gripe cuando nos encontramos estas condiciones.
La semana actual y la anterior poseen	Muy Alta.

frecuencias superiores a la media. La semana actual y la anterior poseen frecuencias superiores a 1.75 veces la de hace 2 semanas.	En el histórico de datos que poseemos, estas condiciones marcan la segunda semana de un brote de gripe.

Revisando la información del año actual, ello nos proporcionará los siguientes datos:

Semana	Frecuencia	Pronóstico
01/12/2013	43	
08/12/2013	37	
15/12/2013	59	
22/12/2013	108	
29/12/2013	309	Inicio del brote de gripe
05/01/2014	420	Segunda semana del brote
12/01/2014	550	
19/01/2014	711	
26/01/2014	468	
02/02/2014	331	
09/02/2014	164	
16/02/2014	88	
23/02/2014	58	
02/03/2014	47	
09/03/2014	51	
16/03/2014	35	
23/03/2014	35	
30/03/2014	33	

Resumiendo, el diseño del cuadro de mandos constará de los siguientes elementos:
-Gráficas con las frecuencias comparativas entre el ICS y la fuente de datos que vamos a utilizar para las predicciones (Google en este caso). Consideramos interesante separar estas gráficas de manera anual, permitiendo al usuario elegir entre una anualidad u otra.

-Datos sobre el mes actual en curso. Preferiblemente en forma de gráfica. Se debe permitir al usuario interactuar con estos datos para obtener información más precisa.

-Información sobre la posibilidad de que exista un brote de gripe, de manera visual.

Por tanto, y teniendo en cuenta todo lo anterior, disponemos los elementos y la información requerida para crear el siguiente diseño del cuadro de mando:

Como podemos ver, en el diseño anterior se ha incluido:
-Gráficos de barras con las frecuencias semanales del ICS y de Google durante la temporada de mayor riesgo de brotes de gripe Esta información está dividida en períodos, los cuales comprenden desde diciembre hasta marzo, y se permite la selección de los dos períodos de los que disponemos información comparativa (2012,2013)

-Componentes y gráficos para seleccionar el mes actual de estudio. Se permite elegir desde diciembre 2013 hasta marzo del 2014. En cada uno de estos períodos se mostrará información sobre las frecuencias semanales ocurridas.

-Información detallada sobre los períodos de estudio e indicación gráfica de la probabilidad de existencia de un brote de gripe.
Seleccionando cada una de estas frecuencias del apartado anterior, se mostrará información numérica adicional (frecuencia actual, frecuencia anterior, media del período), así como la probabilidad de que se produzca un brote de gripe, obteniendo dicha probabilidad en base a los KPI definidos anteriormente.

7 Pruebas realizadas

En este apartado diseñaremos las pruebas necesarias para verificar el buen comportamiento del sistema.
Realizaremos pruebas unitarias de cada una de las fases para comprobar que funcionan correctamente.
Dada la escasa interacción entre las partes, no parece necesario realizar pruebas unitarias, ya que cada una de las fases es independiente del resto.
En este capítulo indicaremos qué y cómo realizar las pruebas. La documentación completa de las mismas, por su extensión, se detallarán en el

7.1 Pruebas de carga.

En esta fase vamos a verificar que los datos se cargan correctamente en la base de datos correspondiente.
Los datos iniciales de la carga se nos proporcionan en hojas excel, y los cargaremos en las tablas TC_ORI de la base de datos.
Revisaremos en este punto:
-Que el número de datos iniciales corresponde con el cargado finalmente en la base de datos.
-Que los datos cargados corresponden con los originales.

Tablas implicadas:

- TC_ORI_GOOGLE
- TC_ORI_URGENCIAS
- TC_ORI_HOSPITAL
- TC_ORI_TWITTER
- TW_WRK_SERVICIO

Hay que hacer notar que la tabla TW_WRK_SERVICIO se carga directamente, por lo que este proceso debe hacerse en la fase de carga. No obstante, dado que es una tabla auxiliar para el resto de tablas de working, se ha adoptado dicha nomenclatura, en lugar de la de las tablas origen (TC_ORI)

7.2 Pruebas de incorporación y procesamiento de los datos.

En esta fase partimos de los datos crudos cargados en la fase anterior, y realizamos una serie de transformaciones hasta llegar a los datos finales.
Existen además una serie de tablas intermedias que servirán para verificar la bondad del proceso.

Comprobaremos los siguientes procesos:

Incorporación de datos TC_ORI_GOOGLE -> TW_WAR_GOOGLE
Este proceso es el más sencillo de todos, ya que es un traspaso simple entre la tabla de origen y la final, sin ningún tipo de filtro.
Verificaremos:
-Que el número de registros entre ambas tablas coinciden.
-Que los campos de datos son los correctos.

Incorporación de los datos del ICS
Este proceso consta de dos partes.
En el primero, cargaremos aquellos datos provenientes de las tablas ORI a las de trabajo.
Las tablas interesadas son:
Entrada: TC_ORI_URGENCIAS, TC_ORI_HOSPITAL
Salida: TW_WRK_ICS, TW_WRK_DIAGNOSTICO

Verificaremos:
-Que el número de registros de TW_WRK_ICS sean la suma de TC_ORI_URGENCIAS y TC_ORI_HOSPITAL, aplicando el filtro correspondiente.
-Que la tabla TW_WRK_DIAGNOSTICO posea el mismo número de registros que TC_ORI_HOSPITAL, aplicando el filtro correspondiente.
-Que los datos cargados en las tablas destino correspondan a los de origen.

Una vez tengamos la tabla intermedia TW_WRK_ICS, obtendremos la tabla final TW_WAR_ICS, agrupando los datos según los períodos de la tabla TC_ORI_GOOGLE.
Comprobaremos:
-Que la suma de frecuencias de TW_WAR_ICS coincide con el número de registros de TW_WRK_ICS.
-Que la agrupación por fechas se ha realizado correctamente.

Incorporación de datos de Twitter.
En este punto, cargaremos los datos de origen de Twitter a las tablas de trabajo.
Las tablas interesadas son:
Entrada: TC_ORI_TWITTER
Salida: TW_WRK_TWITTER, TW_WRK_TUITS

Verificaremos:
-Que el número de registros de TC_ORI_TWITTER sean los mismos que TW_WRK_TWITTER y TW_WRK_TUITS, una vez filtrados por idioma.
-Que los campos de datos son los correctos.

Agrupación por frecuencias de datos de Twitter.

En este proceso, desde la tabla origen TC_ORI_TWITTER, obtendremos la tabla final TW_WAR_TWITTER, agrupando los datos según los períodos de la tabla TC_ORI_GOOGLE.
Comprobaremos:
-Que la suma de frecuencias de TW_WAR_ICS coincide con el número de registros de TW_WRK_ICS.
-Que la agrupación por fechas se ha realizado correctamente.

7.3 Pruebas de informes

Verificaremos que la información obtenida en los informes es correcta.
Para ello, realizaremos unas consultas de la base de datos para verificar que la información se corresponde con la realidad.
Diagnósticos utilizados por fecha y servicio.
Verificaremos contra las tablas TC_WRK_ICS, TW_WRK_DIAGNOSTICO que los datos proporcionados son correctos.

Frecuencia de atenciones por servicio y fecha.
Verificaremos contra la tabla TC_WRK_ICS que los datos proporcionados son correctos. La frecuencia debe coincidir con la existente en TC_WAR_GOOGLE

Listado de tuits por búsqueda e idioma.
Revisaremos contra la tabla TC_ORI_TWITTER que se muestra la información correspondiente por búsqueda e idioma.

Frecuencia comparativa Google-ICS-Twitter.
Para este informe debemos comparar las tablas TC_WRK_ICS, TC_WAR_GOOGLE y TC_WRK_TWITTER. Debemos verificar que las frecuencias de todas las tablas coincidan con la presentada en el gráfico.

7.4 Pruebas del cuadro de mandos.

En el cuadro de mandos poseemos las siguientes informaciones:
-Frecuencias de búsqueda en google por temporada.
-Frecuencias de utilización del ICS por temporada.
-Frecuencias mensuales de búsqueda en google.
-Datos semanales: actual, anterior, media.
-Cálculo de los KPI.
Las pruebas consistirán en verificar que todos estos parámetros son correctos.
Los resultados de todas estas pruebas podrán consultarse en el Anexo 1: Pruebas

8 Pruebas de usabilidad.

En todo proyecto informático es de especial relevancia la participación y supervisión del cliente final desde fases tempranas, para asegurarnos que el producto:
-Satisface una necesidad real.
-Es claro y comprensible.
-Posee una facilidad de uso que permita que el usuario lo adapte como herramienta de trabajo habitual.

Por todo ello, es muy aconsejable contar con una supervisión temprana del trabajo que viene realizándose, incluso con prototipos si es posible antes que con desarrollos terminados.

No obstante, por la casuística de este tipo de proyectos, no es posible contar con la interacción del usuario antes de la entrega del producto final.
Por eso, considero necesario realizar una prueba de usabilidad con una persona ajena al desarrollo, para que sus conclusiones y validaciones puedan suponer un primer filtro para comenzar a pulir ciertos detalles, bien de presentación de la información, bien de manejo, que pudieran mejorar cualitativamente el producto final.

Para ello, vamos a realizar una prueba de usabilidad con usuarios con las siguientes directrices:

8.1 Objetivos del test:

Mediante este test se quiere evaluar la solución aportada (cuadros de mando e informes) en los siguientes aspectos:

a)Eficacia y eficiencia del uso:
-La recuperación de información debe ser rápida, clara y sencilla.
-Una vez comprendidas las herramientas disponibles, el proceso de acceder a los resultados debe ser lo más comprensible y eficaz posible.

b)Contenido útil y práctico:
-Las consultas deben ser claras y sencillas.
-Se deben poder responder a las necesidades del cliente.

c)Claridad de la presentación:
-La información presentada debe hacerse sin ambigüedades
-La presentación debe ser agradable y concisa.

d)Proporción de las tareas:

-Los procesos deben pedir una interacción mínima para comenzar a dar resultados, siendo lo más abreviados posible sin perder claridad.

8.2 Formación pre-test.

Dado que un cuadro de mandos BI y sus informes no es una herramienta ampliamente utilizada por el grueso de los usuarios, es necesaria una primera formación, aunque breve, explicando los distintos informes y cuadro de mandos disponible.
No obstante, dicha formación debe ser breve y sencilla, para no desvirtuar los resultados del test, por lo que consistirá en una somera explicación sobre los informes y controles disponibles, así como la información contenida en la base de datos.
Sin duda, el vídeo explicativo del proyecto será un resumen suficiente para que cualquier usuario pueda comenzar a utilizar la herramienta, y nos remitiremos a él para la formación previa al sujeto de pruebas del test.

8.3 Definición de tareas y escenarios.

Se planifican las siguientes tareas para evaluar la usabilidad de la página.
El usuario debe intentar las tareas que se relatan a continuación.

Escenario	**Tarea**	**Éxito /Fracaso**	**Comentarios**
Consulta de informes	Consultar los diagnósticos que estamos teniendo en cuenta en el informe	Es capaz de consultar el informe	Comentar con el sujeto de pruebas sobre la claridad y presentación de los datos.
Consulta de informes	Comprobar la frecuencia de asistencias por departamento	Es capaz de consultar el informe	Comentar con el sujeto de pruebas sobre la claridad y presentación de los datos.
Manejo del cuadro de mandos	Consulta de las frecuencias google-ics de los distintos períodos	Es capaz de reconocer el mecanismo de selección	
Manejo del cuadro de mandos	Obtener la posibilidad de que exista un brote de gripe la semana del 09/02/2014	Es un éxito el poder obtener el dato requerido	Verificar si el sujeto comprende la información presentada.

8.4 Cuestionario post-test

Una vez terminadas las tareas, se realizará el siguiente cuestionario post-test para evaluar el grado de satisfacción.

Evalúe las siguientes afirmaciones de 1-10 donde 1 significa nada de acuerdo y 10 totalmente de acuerdo.

-Los informes y cuadro de mando son sencillos de utilizar.
-He podido realizar las tareas encomendadas de manera fácil e intuitiva.
-En todo momento sabía el estado de la tarea que estaba realizando.
-En caso de error, he tenido la información suficiente para corregirlo.
-La información se me ha mostrado de manera clara y sencilla.
-Los formularios son cortos y fáciles de entender.
-La experiencia ha sido cómoda, y el proceso, rápido.

Los resultados se informarán en el Anexo 2: Pruebas de usabilidad.

9 Inclusión de datos de redes sociales.

En el estudio se han tomado en cuenta, además de la información facilitada por el ICS, los datos provenientes de dos fuentes: el servicio Google Flu, de google, (http://www.google.org/flutrends/es/), y la información que se ha podido obtener, en años anteriores, de la red social Twitter.
-Datos de google: Obtenidos de http://www.google.org/flutrends/es/data.txt.
Se trata de una serie histórica de datos con información semanal, en la cual se nos indica, por regiones, el número de consultas relacionadas con la gripe en dicha semana.
No es posible modificar los parámetros de búsqueda, debemos confiar en el criterio seguido por Google para recuperar esta información, pero por otra parte confiamos en la fiabilidad de los datos.
Además, es importante reseñar que poseemos información histórica desde el año 2003, lo cual podemos utilizar en nuestro beneficio para conseguir una mayor fiabilidad del proyecto.

-Datos de Twitter. Se han incorporado mediante un fichero obtenido de años anteriores, con el valor de búsqueda "gripe". La información obtenida se refiere al invierno del año 2012 (noviembre 2012 – marzo 2013). Tampoco es posible obtener o filtrar información adicional de esta fuente, ya que se ha incorporado conforme se ha suministrado.

¿Qué otras fuentes podrían utilizarse en este proyecto?
El proyecto está preparado para añadir nuevas fuentes de datos que afiancen los resultados obtenidos, y permita su viabilidad en el tiempo.
Se han incorporado procesos para el tratamiento tanto de frecuencias semanales(como el fichero de google) como diarias (fichero de twitter), por lo que es posible añadir nuevas fuentes de información con un coste de programación mínimo.
Estas fuentes pueden ser:
-Futuras actualizaciones del fichero de google. En un principio, parece que el proyecto de Google Flu va a poseer una continuación en el tiempo que nos permite contar con estos datos para mantener la viabilidad del proyecto.

-Datos descargados de Twitter.
Para ello, recomendaríamos implementar un programa que utilice la API de twitter para ir descargando datos relacionados. Si ampliamos la cantidad de datos de esta fuente, mejoraremos la fiabilidad del proyecto, al contar con dos fuentes independientes de información que nos permitan predecir con mayor exactitud los brotes de gripe que puedan afectar al sistema.
Para futuras incorporaciones de datos, hay que reseñar que quizás no sea adecuado limitarse a buscar tan sólo "gripe", sino que habría que incorporar más filtros de búsqueda como "enfermo", "tosiendo", o términos similares que pudiesen estar relacionados.

Sería interesante realizar una descargas de tuits con un gran número de término y después aplicar técnicas estadísticas (como las realizadas en el apartado Selección de los departamentos del ICS sensibles.) para comprobar cuáles son los términos que realmente nos aportan la información que estamos buscando.

-Otras redes sociales.
Durante el año 2013 se han incorporado #hashtags a otras redes sociales, como Facebook y LinkedIn. Por ejemplo, podemos recuperar información relacionada con la gripe en la red de Facebook con el link :
https://www.facebook.com/hashtag/gripe?fref=ts
No obstante, no es posible aún recuperar esta información con la API de Facebook aún. Parece ser que existe una primera versión de esta funcionalidad, pero limitada a un número cerrado de desarrolladores:
https://developers.facebook.com/docs/public_feed/

En el momento en el que la API pública permita dichas funcionalidades, sería posible implementar un programa similar al planteado con Twitter para recuperar información de esta red también.

Más información sobre la API de Facebook:
https://developers.facebook.com/docs/apps

10 Conclusiones

Tras el trabajo elaborado, es en este capítulo cuando se puede echar la vista atrás y plasmar las conclusiones y percepciones de tipo personal que se han estado formando durante el proyecto.

Una de las primeras conclusiones que me viene a la cabeza es la propia mutabilidad de la red. Lo que es posible un año, puede que al que viene sea mucho más asequible, con mejores herramientas, o, en casos bastante raros, que la forma de realizarlo cambie totalmente.
Uno de los primeros tropiezos que ha tenido este proyecto es la recuperación de datos de twitter. A pesar de tener una api bien documentada, y ampliamente utilizada, en su última versión ha incorporado numerosos cambios, sobre todo enfocados a la seguridad, que han cambiado la manera de programar con ella en gran medida.
La primera de las medidas adoptadas ha sido que no es posible recuperar información histórica que no pertenezca al propio perfil con una antigüedad superior a 7 días.
Ello ha hecho inviable la realización de este proyecto tal y como se había planteado, pero afortunadamente, la red proporciona un buen número de alternativas, y el abanico que se puede utilizar es amplio.
La propia empresa google ya ha intentado realizar un muestreo de la frecuencia en la que se busca en sus servidores la palabra “gripe”, y esta alternativa es la que hemos utilizado para este proyecto.
No obstante, y saliéndonos del ámbito de este proyecto en concreto, algunas redes sociales comienzan a implementar un mecanismo de hashtags, e incluyen las correspondientes Apis para tratar de recuperar esta información, por lo que podría ser que en breve podamos utilizar facebook como fuente de información para este proyecto, o bien realizar estudios similares en LinkedIn para recuperar y tratar información que pudiera ser interesante como método de estudio.

Haciendo un inciso en este apartado, quisiera reseñar la cantidad de información que aportamos voluntariamente a numerosas empresas privadas.
Hoy en día llevamos nuestros dispositivos móviles a todas partes, estamos conectados permanentemente a internet, y nos gusta compartir con nuestros amigos y conocidos información que nos es relevante, o bien simplemente compartir por el simple placer de hacerlo, y mantener una especie de relación virtual con otras personas.
¿A dónde va esta información? ¿Quién la guarda?
Porque, del mismo modo que google guarda información de con qué frecuencia se busca la palabra “gripe”, no debemos pensar que no lo están haciendo con otros términos.
Quizás en plena campaña de navidad, google realiza estudios de mercado sobre lo que estamos buscando, y venda esta información con jugosos beneficios a importantes compañías.

Por lo pronto, podemos estar seguros de que si buscamos un cierto producto en Amazon, google, inmediatamente y al navegar por otras páginas veremos que hay banners sobre artículos relacionados con nuestra búsqueda anterior, y no es casualidad.
¿Qué huella estamos dejando a nuestro paso? ¿Quién está en conocimiento de esta información?
Esta es la era del Big Data, y ello está cambiando nuestra vida más de lo que pensamos.
Una muestra de ello es este proyecto, que utiliza información de gente que la ha compartido libremente, pero sin conocimiento alguno de que nosotros íbamos a utilizarla de esta manera.

Hay una frase que podría resumir lo que está pasando, y que debemos tener en cuenta cada vez que utilizamos un servicio de calidad y , en apariencia gratuito : “Si no sabes cuál es el negocio, quizás el negocio seas tú”. Con ello queremos hacer pensar que nada es gratis, todo tiene un coste, y que google, facebook, twitter, etc... no son servicios gratuitos, sino que se financian en base a la información que nosotros les aportamos.

Otro de los temas de los que he tomado conciencia durante el proyecto es de la verdadera naturaleza del open source.
Al utilizar Pentaho, podemos utilizar una versión comercial u otra open source.
Sin duda, la alternativa más económica es la open source (gratuita), frente a los costes, normalmente bastante altos, que puede suponer un programa comercial.
Los primeros pasos en un software siempre son traumáticos.
Y si no se dispone de documentación completa y actualizada, seguro que lo son más aún.
Al intentar progresar en el proyecto con un software open source, me he topado con bastantes barreras. La documentación era, en el mejor de los casos, desactualizada. Tras numerosas búsquedas en el foro, llegué a encontrar un manual completo en vídeo de la última versión que, tras intentar acceder a él, comprobé con desconcierto que el servidor estaba caído desde hace un mes.
La información que pude encontrar en el foro fue de bastante ayuda, pero no poseía una centralización correcta, ni una rápida búsqueda.
Además, la ausencia de una ayuda en línea en el propio programa retrasa mucho la realización de las tareas más básicas.
No es sencillo programar con este tipo de programas, pero al mismo tiempo he de decir que no es imposible.
Tras la impotencia inicial, y el incremento exponencial de los tiempos que había reservado a formación, poco a poco la situación se normalizó, y pude comprobar que, aunque a veces caótico y sin duda mal documentado, la verdad es que el software open source funciona, y una vez terminado el proyecto poseía un programa robusto y confiable.
¿Es adecuado a todas las empresas un programa de estas características? Es muy difícil que una empresa mediana pueda afrontar una plataforma de estas características con el apoyo tan sólo de las colaboraciones interesadas de otros usuarios, sobre todo cuando existen necesidades muy concretas y plazos

ajustados para realizar una labor de cara a otras empresas. No obstante, un profesional bien entrenado en esta labor, sin duda supondría un importantísimo ahorro de costes. Hay que tener en cuenta que, por ejemplo, la licencia de qlick para un servidor cuesta 35.000 usd (http://www.qlik.com/es/explore/pricing)

En cuanto a la planificación del proyecto, y por los motivos anteriores, he comprobado como los plazos de aquellas herramientas o procesos que conocía (análisis, diseño, documentación, pruebas, código SQL) ha sido cumplido según lo previsto, con un porcentaje de desviación relativamente bajo respecto a los estudios iniciales.
No obstante, he sido demasiado optimista en la generación de código con herramientas en las que no poseía experiencia.
Sin duda, este caso puede extrapolarse a cualquier proyecto formal: la inclusión de herramientas en las que no poseemos experiencia pueden suponer importantes desfases en costes y plazos que pueden ser difíciles de asumir.
Sin duda, estos problemas podrían paliarse realizándose un examen más minucioso del funcionamiento de nuevas herramientas, incluyendo una fase de aprendizaje previo más exhaustiva, y en el caso de proyectos con varias personas implicadas, en la contratación o inclusión en el equipo de un experto que pueda ayudar en aquellos puntos que revisten de una mayor dificultad, y que para el experto bien entrenado pueden ser resueltos en minutos, frente a jornadas enteras para aquellos programadores que no tienen experiencia en esta herramienta en concreto.

En estos momentos, podemos concretar que el proyecto está completado, pero sería mucho decir que estuviera finalizado.
Un programa que no se actualiza, amplía, mejora, sin duda está muerto por definición, y ahora mismo lo más interesante sería presentar la programación al usuario final para que éste comience a proponer mejoras o funcionalidades que no han sido contempladas por el programador. Sin duda, los requerimientos que puede pensar un informático son a veces diametralmente opuestos a aquellos que puede necesitar el usuario para el día a día.
Además la precisión del cuadro de mandos será correcta siempre y cuando los datos sigan estando actualizados. Y para ello, no sólo tenemos que contar con datos del ICS, sino que podemos incorporar nuevas redes sociales para mejorar las predicciones realizadas.
Un seguimiento semanal de la red Twitter nos podría proporcionar nuevos datos que utilizar en próximas temporadas, e incluso, dada la frecuencia de actualización de esta fuente, podríamos llegar a establecer predicciones diarias en lugar de semanales. Y, cuando se permita el acceso a los nuevos hashtags de facebook mediante su api, podríamos incluir también esta red social para ampliar la base de datos disponible.
Todo ello redundaría en una mayor precisión y rapidez de detección de la aplicación.

11 Glosario

Api : Interfaz de programación de aplicaciones (IPA) o API (del inglés Application Programming Interface) es el conjunto de funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción. Son usadas generalmente en las bibliotecas.

BI: (Del inglés business intelligence) Se denomina inteligencia empresarial, inteligencia de negocios o BI al conjunto de estrategias y aspectos relevantes enfocadas a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.

Big Data: Es, en el sector de tecnologías de la información y la comunicación, una referencia a los sistemas que manipulan grandes conjuntos de datos (o data sets). Las dificultades más habituales en estos casos se centran en la captura, el almacenamiento, búsqueda, compartición, análisis, y visualización.

Hashtag: (del inglés hash, almohadilla o numeral y tag, etiqueta)2 es una cadena de caracteres formada por una o varias palabras concatenadas y precedidas por una almohadilla o gato (#). Es, por lo tanto, una etiqueta de metadatos precedida de un carácter especial con el fin de que tanto el sistema como el usuario la identifiquen de forma rápida.

ICS: Instituto Catalán de Salud.

KPI : del inglés Key Performance Indicators, o Indicadores Clave de Desempeño, son métricas utilizadas para medir miden el nivel del desempeño de un proceso, centrándose en el "cómo" e indicando el rendimiento de los procesos, de forma que se pueda alcanzar el objetivo fijado.

LOPD: Ley orgánica de protección de datos. Ley española que se dedica a la protección de datos personales. Ver Material de apoyo utilizado

Tuit: Dícese de cada uno de los mensajes publicados en la red Twitter.

Open source: Código abierto es la expresión con la que se conoce al software distribuido y desarrollado libremente. Se focaliza más en los beneficios prácticos (acceso al código fuente) que en cuestiones éticas o de libertad que tanto se destacan en el software libre.

12 Material de apoyo utilizado

Durante el desarrollo del proyecto se ha utilizado información recogida de distintas fuentes. Entre ellas podemos destacar:

Materiales utilizados:
http://www.20minutos.es/noticia/2089892/0/domotica/hogar/internet/
Gráfico del apartado 1
Artículo sobre la evolución de los dispositivos conectados a internet, y el "internet de las cosas".

Definiciones y material de consulta:
http://es.wikipedia.org/

Material de la asignatura:
Introducción al Business Intelligence
Jordi Conesa Caralt (coord.)
Josep Curto Díaz
ISBN: 978-84-9788-979-7

Material sobre Pentaho:
http://forums.pentaho.com/
Foro de desarrolladores de Pentaho.

http://pentahohispano.blogspot.com.es/2011/07/como-hacer-cuadros-de-mando-v.html
Apuntes sobre cuadros de mando

Texto LOPD:
http://www.agpd.es/portalwebAGPD/canaldocumentacion/legislacion/estatal/common/pdfs/LOPD_consolidada.pdf
Ley orgánica de protección de datos personales.

13 Anexo 1: Pruebas

Incluimos en este anexo las pruebas descritas en el apartado Pruebas realizadas
Se intentará validar la bondad de la programación realizada.

13.1Pruebas de carga

Se nos proporcionan 3 hojas excel como datos de partida.
BI_UOC_Datos_Hospitalizacionv1.xls : 14326 registros.
CasBI_UOC_v1Urgencias.xlsx 50518 registros.
tweets_gripe_2013.csv: 2101 registros.

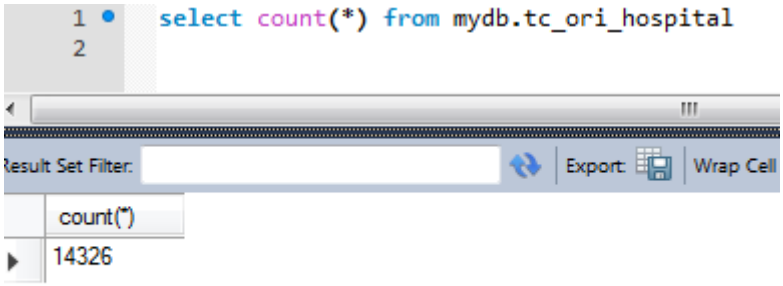
Adicionalmente, obtenemos la información de Google Flu:
http://www.google.org/flutrends/es/data.txt : 548 registros.

Y cargaremos la tabla TW_WRK_SERVICIO con un SQL que inserta 50 elementos.

Tras lanzar el proceso de carga, realizamos las siguientes pruebas.

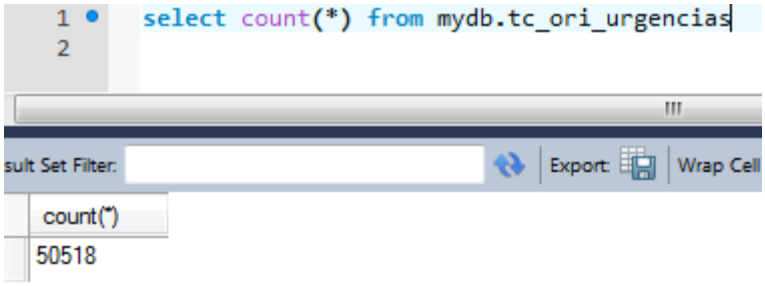
Número de registros incorporados en las tablas:

- TC_ORI_HOSPITAL



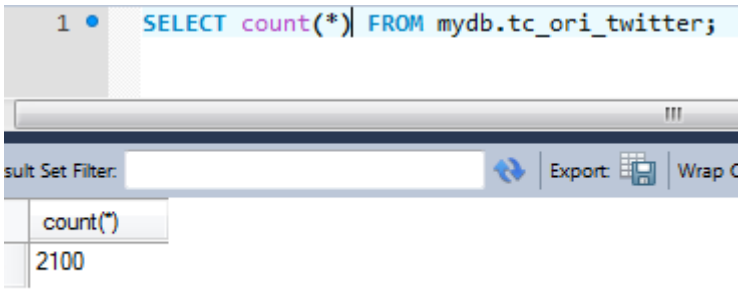
Comprobamos que el número de registros de la tabla es igual al esperado.

- TC_ORI_URGENCIAS



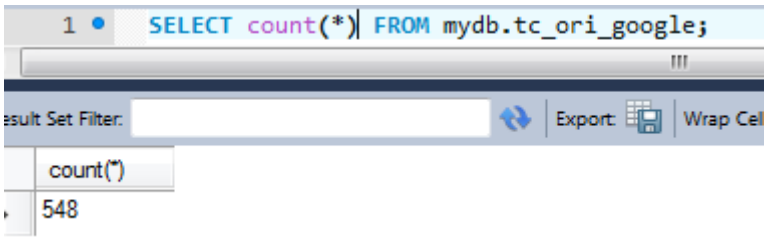
Comprobamos que el número de registros de la tabla es igual al esperado.

- TC_ORI_TWITTER



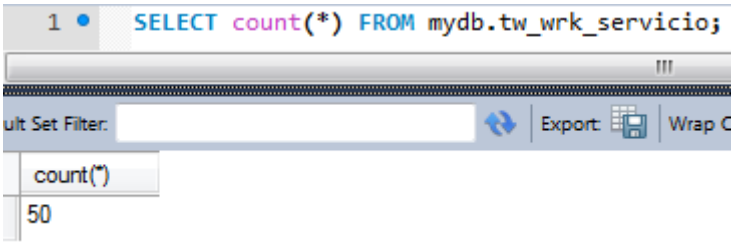
Comprobamos que el número de registros de la tabla es igual al esperado.

- TC_ORI_GOOGLE



Comprobamos que el número de registros de la tabla es igual al esperado.

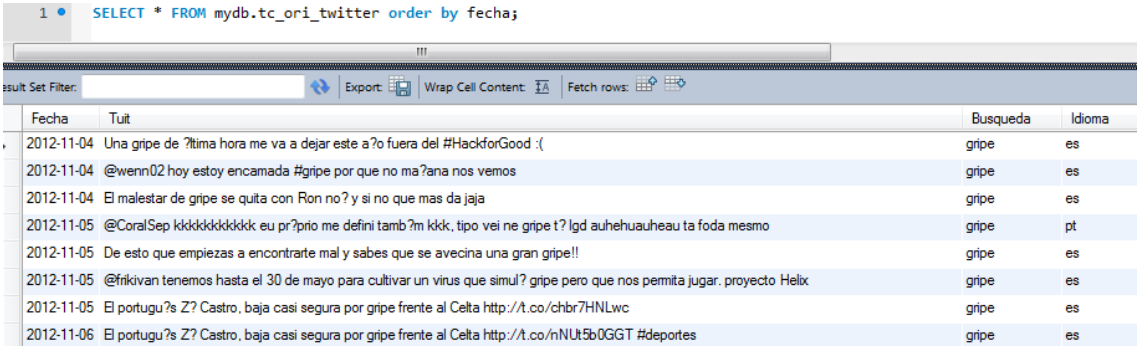
- TC_WRK_SERVICIO



Comprobamos que el número de registros de la tabla es igual al esperado.

Revisamos que la información incorporada corresponde a la esperada:

- TC_ORI_TWITTER



Fecha	Tuit	Busqueda	Idioma
2012-11-04	Una gripe de ?ltima hora me va a dejar este a?o fuera del #HackforGood :(	gripe	es
2012-11-04	@wenn02 hoy estoy encamada #gripe por que no ma?ana nos vemos	gripe	es
2012-11-04	El malestar de gripe se quita con Ron no? y si no que mas da jaja	gripe	es
2012-11-05	@CoralSep kkkkkkkkkkk eu pr?prio me defini tamb?m kkk, tipo vei ne gripe t? lgd auhehuauheau ta foda mesmo	gripe	pt
2012-11-05	De esto que empiezas a encontrarte mal y sabes que se avecina una gran gripe!!	gripe	es
2012-11-05	@frikivan tenemos hasta el 30 de mayo para cultivar un virus que simul? gripe pero que nos permita jugar. proyecto Helix	gripe	es
2012-11-05	El portugu?s Z? Castro, baja casi segura por gripe frente al Celta http://t.co/chbr7HNLwc	gripe	es
2012-11-06	El portugu?s Z? Castro, baja casi segura por gripe frente al Celta http://t.co/nNUt5b0GGT #deportes	gripe	es

1	Una gripe de última hora me va a dejar este año fuera del #HackforGood :(	gripe	04/11/2012	es
2	@wenn02 hoy estoy encamada #gripe por que no mañana nos vemos	gripe	04/11/2012	es
3	El malestar de gripe se quita con Ron no? y si no que mas da jaja	gripe	04/11/2012	es
4	@CoralSep kkkkkkkkkkk eu próprio me defini também kkk, tipo vei ne gripe tá lgd auhehuauheau ta foda mesmo	gripe	05/11/2012	pt
5	De esto que empiezas a encontrarte mal y sabes que se avecina una gran gripe!!	gripe	05/11/2012	es
6	@frikivan tenemos hasta el 30 de mayo para cultivar un virus que simulé gripe pero que nos permita jugar. proyecto Helix	gripe	05/11/2012	es
7	El portugués Zé Castro, baja casi segura por gripe frente al Celta http://t.co/chbr7HNLwc	gripe	05/11/2012	es
8	El portugués Zé Castro, baja casi segura por gripe frente al Celta http://t.co/nNUt5b0GGT #deportes	gripe	06/11/2012	es

Verificamos que la información cargada corresponde con la existente en la hoja excel.

- TC_ORI_HOSPITAL

Fecha_Ingreso	Servicio	Descripcion_diagnostico
2008-06-18	CGDSVGT	Hematoma que complica un procediment
2011-06-12	CGDSVGT	Enteritis regional intesti prim. Malaltia Crohn de:duode,ili,jeju,ileitis:regional, segmentaria,terminal
2011-09-14	CGDSVGT	Atencio d'ileostomia
2011-10-15	NMLSVGT	Actinomicosi pulmonar. Actinomicosi toracica
2011-10-25	MIVSVGT	Aterosclerosi d'arteria coronaria nadiua
2011-11-07	MIRSVGT	Pneumonia bacteriana inespecificada
2011-11-08	HADUFGT	Pielonefritis aguda sense lesio de necrosi medul.lar renal
2011-11-09	HADUFGT	Diabetis amb trastoms circulatoris periferics, tipus II controlada

1	18/06/2008	CGDSVGT	Hematoma que complica un procediment
2	12/06/2011	CGDSVGT	Enteritis regional intesti prim. Malaltia Crohn de:duode,ili,jeju,ileitis:regional, segmentaria,terminal
3	14/09/2011	CGDSVGT	Atencio d'ileostomia
4	15/10/2011	NMLSVGT	Actinomicosi pulmonar. Actinomicosi toracica
5	25/10/2011	MIVSVGT	Aterosclerosi d'arteria coronaria nadiua
6	07/11/2011	MIRSVGT	Pneumonia bacteriana inespecificada
7	08/11/2011	HADUFGT	Pielonefritis aguda sense lesio de necrosi medul.lar renal
8	09/11/2011	HADUFGT	Diabetis amb trastorns circulatoris periferics, tipus II controlada

Verificamos que la información cargada corresponde con la existente en la hoja excel.

- TC_ORI_URGENCIAS

```
1 • SELECT * FROM mydb.tc_ori_urgencias order by fecha_entrada
```

ult Set Filter: Export: Wrap Cell Content: Fetch

Fecha_Entrada	Servicio_Entrada	Servicio_Alta
2011-12-22	COTSVGT	COTSVGT
2011-12-30	CGDSVGT	CGDSVGT
2011-12-30	ACVSVGT	COTSVGT
2011-12-30	UROSVGT	UROSVGT

1	22/12/2011	COTSVGT	COTSVGT
2	30/12/2011	UROSVGT	UROSVGT
3	30/12/2011	ACVSVGT	COTSVGT
4	30/12/2011	CGDSVGT	CGDSVGT

Verificamos que la información cargada corresponde con la existente en la hoja excel.

- TC_ORI_GOOGLE

```
1 •  SELECT * FROM mydb.tc_ori_google order by fecha_intervalo
```

esult Set Filter: Export: Wrap Cell Content:

Fecha_Intervalo	Frecuencia
2003-10-05	0
2003-10-12	35
2003-10-19	66
2003-10-26	105
2003-11-02	137
2003-11-09	107
2003-11-16	353
2003-11-23	130

1	05/10/2003	0
2	12/10/2003	35
3	19/10/2003	66
4	26/10/2003	105
5	02/11/2003	137
6	09/11/2003	107
7	16/11/2003	353
8	23/11/2003	130

Verificamos que la información cargada corresponde con la existente en la hoja excel.

- TW_WRK_SERVICIO

Id_Servicio	Nombre_Servicio	Descripcion_Servicio
1	ACVSVGT	ANGIOLOGIA I CIRURGIA VASCULAR
2	ADIUFGT	UNITAT ADDICCIONS - DESINTOXIC
3	ANCSVGT	SIN DESCRIPCION DISPONIBLE
4	ANESVGT	ANESTESIOLOGIA I REANIMACIO
5	APASVGT	SIN DESCRIPCION DISPONIBLE
6	CARSVGT	CARDIOLOGIA
7	CCASVGT	CIRURGIA CARDIACA
8	CGDSVGT	CIRURGIA GENERAL I DIGESTIVA

```
INSERT INTO mydb.tw_wrk_servicio VALUES (01,"ACVSVGT","ANGIOLOGIA I CIRURGIA VASCULAR");
INSERT INTO mydb.tw_wrk_servicio VALUES (02,"ADIUFGT","UNITAT ADDICCIONS - DESINTOXIC");
INSERT INTO mydb.tw_wrk_servicio VALUES (03,"ANCSVGT","SIN DESCRIPCION DISPONIBLE    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (04,"ANESVGT","ANESTESIOLOGIA I REANIMACIO   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (05,"APASVGT","SIN DESCRIPCION DISPONIBLE    ");
INSERT INTO mydb.tw_wrk_servicio VALUES (06,"CARSVGT","CARDIOLOGIA                   ");
INSERT INTO mydb.tw_wrk_servicio VALUES (07,"CCASVGT","CIRURGIA CARDIACA             ");
INSERT INTO mydb.tw_wrk_servicio VALUES (08,"CGDSVGT","CIRURGIA GENERAL I DIGESTIVA  ");
```

Verificamos que la información cargada corresponde con el SQL creado para su carga.

Con estas pruebas hemos comprobado que el proceso de carga es correcto.

13.2 Pruebas de incorporación y procesamiento de los datos.

En esta fase partimos de los datos crudos cargados en la fase anterior, y realizamos una serie de transformaciones hasta llegar a los datos finales.
Revisaremos en estas pruebas tanto las tablas intermedias como las finales.

Comprobaremos los siguientes procesos:

Incorporación de datos TC_ORI_GOOGLE -> TW_WAR_GOOGLE
Consiste en un traspaso simple entre la tabla de origen y la final, sin ningún tipo de filtro.
Verificaremos:
-Que el número de registros entre ambas tablas coinciden.

```
1  SELECT 'tc_ori_google', count(*)  FROM mydb.tc_ori_google
2  union
3  SELECT 'tw_war_google', count(*)  FROM mydb.tw_war_google
```

esult Set Filter: Export: Wrap Cell Content:

tc_ori_google	count(*)
tc_ori_google	548
tw_war_google	548

-Que los campos de datos son iguales.

```
1  select * from
2  (SELECT 'tc_ori_google', a.*   FROM mydb.tc_ori_google a where a.fecha_intervalo < '2003-11-09'
3  union
4  SELECT 'tw_war_google', b.*  FROM mydb.tw_war_google b where b.fecha_intervalo < '2003-11-09') c
5  order by c.fecha_intervalo
```

sult Set Filter: Export: Wrap Cell Content:

tc_ori_google	Fecha_Intervalo	Frecuencia
tc_ori_google	2003-10-05	0
tw_war_google	2003-10-05	0
tc_ori_google	2003-10-12	35
tw_war_google	2003-10-12	35
tc_ori_google	2003-10-19	66
tw_war_google	2003-10-19	66
tw_war_google	2003-10-26	105
tc_ori_google	2003-10-26	105
tc_ori_google	2003-11-02	137
tw_war_google	2003-11-02	137

Incorporación de los datos del ICS

Este proceso consta de dos partes.

En el primero, cargaremos aquellos datos provenientes de las tablas ORI a las de trabajo.

Las tablas interesadas son:

Entrada: TC_ORI_URGENCIAS, TC_ORI_HOSPITAL

Salida: TW_WRK_ICS, TW_WRK_DIAGNOSTICO

Verificaremos:

-Que el número de registros de TW_WRK_ICS sean la suma de TC_ORI_URGENCIAS y TC_ORI_HOSPITAL, aplicando el filtro correspondiente.

```
select * from
(SELECT 'tc_ori_hospital' as tabla, count(*)   FROM mydb.tc_ori_hospital a
where a.servicio in ('PEESVGT','URGSVGT','GINSCGT')
union
SELECT 'tc_ori_urgencias'as tabla, count(*)  FROM mydb.tc_ori_urgencias b
where (b.servicio_entrada in ('PEESVGT','URGSVGT','GINSCGT') or
b.servicio_alta in ('PEESVGT','URGSVGT','GINSCGT'))
union
SELECT 'tw_wrk_ics'as tabla, count(*)  FROM mydb.tw_wrk_ics c
) d
```

esult Set Filter: Export: Wrap Cell Content:

tabla	count(*)
tc_ori_hospital	1853
tc_ori_urgencias	32364
tw_wrk_ics	34217

-Que la tabla TW_WRK_DIAGNOSTICO posea el mismo número de registros que TC_ORI_HOSPITAL, aplicando el filtro correspondiente.

```
select * from
(SELECT 'tc_ori_hospital' as tabla, count(*)   FROM mydb.tc_ori_hospital a
where a.servicio in ('PEESVGT','URGSVGT','GINSCGT')
union
SELECT 'tw_wrk_diagnostico'as tabla, count(*)  FROM mydb.tw_wrk_diagnostico b
) d
```

ult Set Filter: Export: Wrap Cell Content:

tabla	count(*)
tc_ori_hospital	1853
tw_wrk_diagnostico	1853

-Que los datos cargados en las tablas destino correspondan a los de origen.
Datos de urgencias:

```
1 • SELECT * FROM mydb.tw_wrk_ics where Id_Origen = 1 and fecha_ingreso = '2011-12-31';
```

Result Set Filter: Export: Wrap Cell Content:

Id_Origen	Fecha_Ingreso	Id_Servicio_Entrada	Id_Servicio_Salida	Id_Diagnostico
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	34	34	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL
1	2011-12-31	49	49	NULL

*Datos de los servicios implicados

```
1 • SELECT * FROM mydb.tw_wrk_servicio where id_servicio in (34,49)
```

Result Set Filter: Edit: Export/Import: Wrap

Id_Servicio	Nombre_Servicio	Descripcion_Servicio
34	PEESVGT	PEDIATRIA
49	URGSVGT	MEDICINA D'URGENCIES

```
1 • SELECT * FROM mydb.tc_ori_urgencias where fecha_entrada = '2011-12-31'
2   and (servicio_entrada in ('PEESVGT','URGSVGT','GINSCGT') or
3   servicio_alta in ('PEESVGT','URGSVGT','GINSCGT') );
```

Result Set Filter: Export: Wrap Cell Content:

Fecha_Entrada	Servicio_Entrada	Servicio_Alta
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	PEESVGT	PEESVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT
2011-12-31	URGSVGT	URGSVGT

Datos de hospitalización:

```
SELECT * FROM mydb.tw_wrk_ics where Id_Origen = 2 and fecha_ingreso = '2011-12-31';
```

Export: Wrap Cell Content:

Id_Origen	Fecha_Ingreso	Id_Servicio_Entrada	Id_Servicio_Salida	Id_Diagnostico
2	2011-12-31	49	NULL	4
2	2011-12-31	34	NULL	5
2	2011-12-31	34	NULL	6
2	2011-12-31	49	NULL	237
2	2011-12-31	49	NULL	261
2	2011-12-31	49	NULL	262
2	2011-12-31	34	NULL	263

```
SELECT *   FROM mydb.tc_ori_hospital a
where a.servicio in ('PEESVGT','URGSVGT','GINSCGT')
and a.Fecha_Ingreso = '2011-12-31'
```

Export: Wrap Cell Content:

Fecha_Ingreso	Servicio	Descripcion_diagnostico
2011-12-31	URGSVGT	Infeccio vies urinaries, localitzacio no especificada. Bacteriuria, piuria
2011-12-31	PEESVGT	Altres infeccions especificades del periode perinatal. Infec.nado NOS,infec.intraamnica fetus NOS
2011-12-31	PEESVGT	Epilepsia convulsiva generalitzada, sense mencio d'epilepsia intractable
2011-12-31	URGSVGT	Bronquitis aguda. Bronquitis aguda/subaguda:fibrinosa,membranosa,purulenta,septica,virica,bronq.crupal
2011-12-31	URGSVGT	Bronquitis aguda. Bronquitis aguda/subaguda:fibrinosa,membranosa,purulenta,septica,virica,bronq.crupal
2011-12-31	URGSVGT	Fibril.lacio auricular
2011-12-31	PEESVGT	Altres infeccions especificades del periode perinatal. Infec.nado NOS,infec.intraamnica fetus NOS

```
SELECT * FROM mydb.tw_wrk_diagnostico
where id_diagnostico in (4,5,6,237,261,262,263)
```

Edit: Export/Import: Wrap Cell Content:

Id_Diagnostico	Descripcion_Diagnostico
4	Infeccio vies urinaries, localitzacio no especificada. Bacteriuria, piuria
5	Altres infeccions especificades del periode perinatal. Infec.nado NOS,infec.intraamnica fetus NOS
6	Epilepsia convulsiva generalitzada, sense mencio d'epilepsia intractable
237	Bronquitis aguda. Bronquitis aguda/subaguda:fibrinosa,membranosa,purulenta,septica,virica,bronq.crupal
261	Bronquitis aguda. Bronquitis aguda/subaguda:fibrinosa,membranosa,purulenta,septica,virica,bronq.crupal
262	Fibril.lacio auricular
263	Altres infeccions especificades del periode perinatal. Infec.nado NOS,infec.intraamnica fetus NOS

Hemos comprobado que los datos originales se han cargado correctamente en el resto de tablas (TW_WRK_DIAGNOSTICO, TW_WRK_ICS)

Verificamos la incorporación de datos a la tabla TC_WAR_ICS
Esta tabla es una agrupación de la tabla intermedia TC_WRK_ICS, según los períodos de la tabla TC_ORI_GOOGLE.

Comprobaremos:
-Que la suma de frecuencias de TW_WAR_ICS coincide con el número de registros de TW_WRK_ICS.

```
1  Select tabla, frecuencia from
2  (SELECT 'tw_war_ics' as tabla, sum(frecuencia) as Frecuencia FROM mydb.tw_war_ics
3  union
4  SELECT 'tw_wrk_ics' as tabla, count(*) as Frecuencia FROM mydb.tw_wrk_ics) c
```

ult Set Filter: Export: Wrap Cell Content:

tabla	frecuencia
tw_war_ics	34217
tw_wrk_ics	34217

-Que la agrupación por fechas se ha realizado correctamente.

```
1  SELECT * FROM mydb.tw_war_ics where Fecha_intervalo = '2012-12-23'
```

sult Set Filter: Export: Wrap Cell Content:

Fecha_Intervalo	Frecuencia
2012-12-23	18

```
1  SELECT count(*) FROM mydb.tw_wrk_ics where (Fecha_ingreso >= '2012-12-23'
2  and Fecha_ingreso < '2012-12-30')
```

ult Set Filter: Export: Wrap Cell Content:

count(*)
18

Incorporación de datos de Twitter.
En este punto, cargaremos los datos de origen de Twitter a las tablas de trabajo.
Las tablas interesadas son:
Entrada: TC_ORI_TWITTER
Salida: TW_WRK_TWITTER, TW_WRK_TUITS

Verificaremos:

-Que el número de registros de TC_ORI_TWITTER sean los mismos que TW_WRK_TWITTER y TW_WRK_TUITS, una vez filtrados por idioma.

```
1  Select tabla, frecuencia from
2  (SELECT 'tc_ori_twitter' as tabla, count(*) as Frecuencia FROM mydb.tc_ori_twitter
3  where idioma = 'es'
4  union
5  SELECT 'tw_wrk_twitter' as tabla, count(*) as Frecuencia FROM mydb.tw_wrk_twitter) c
```

sult Set Filter: Export: Wrap Cell Content:

tabla	frecuencia
tc_ori_twitter	1903
tw_wrk_twitter	1903

-Que los campos de datos son los correctos.

```
1 •  SELECT * FROM mydb.tc_ori_twitter
2    where fecha = '2012-11-04'
```

ult Set Filter: Export: Wrap Cell Content:

Fecha	Tuit	Busqueda	Idioma
2012-11-04	Una gripe de ?ltima hora me va a dejar este a?o fuera del #HackforGood :(	gripe	es
2012-11-04	@wenn02 hoy estoy encamada #gripe por que no ma?ana nos vemos	gripe	es
2012-11-04	El malestar de gripe se quita con Ron no? y si no que mas da jaja	gripe	es

```
1 •  SELECT * FROM mydb.tw_wrk_twitter
2    where fecha_tuit = '2012-11-04'
```

lt Set Filter: Export:

Fecha_Tuit	ID_Tuit	Busqueda	Idioma
2012-11-04	1	gripe	es
2012-11-04	2	gripe	es
2012-11-04	3	gripe	es

```
1 •  SELECT * FROM mydb.tw_wrk_tuits
2    where id_tuit in (1,2,3)
```

ult Set Filter: Edit: Export/Import:

Id_Tuit	Tuit
1	Una gripe de ?ltima hora me va a dejar este a?o fuera del #HackforGood :(
2	@wenn02 hoy estoy encamada #gripe por que no ma?ana nos vemos
3	El malestar de gripe se quita con Ron no? y si no que mas da jaja

Agrupación por frecuencias de datos de Twitter.

En este proceso, desde la tabla TW_WRK_TWITTER, obtendremos la tabla final TC_WAR_TWITTER, agrupando los datos según los períodos de la tabla TC_ORI_GOOGLE.
Comprobaremos:
-Que la suma de frecuencias de TW_WAR_ICS coincide con el número de registros de TW_WRK_ICS.

```
1  Select tabla, frecuencia from
2  (SELECT 'tw_war_ics' as tabla,sum(frecuencia) as Frecuencia FROM mydb.tw_war_ics
3  union
4  SELECT 'tw_wrk_ics' as tabla, count(*) as Frecuencia FROM mydb.tw_wrk_ics) c
```

sult Set Filter: Export: Wrap Cell Content:

tabla	frecuencia
tw_war_ics	34217
tw_wrk_ics	34217

-Que la agrupación por fechas se ha realizado correctamente.

```
1  SELECT * FROM mydb.tw_war_ics where fecha_intervalo = '2011-12-25'
```

lt Set Filter: Export: Wrap Cell Content:

Fecha_Intervalo	Frecuencia
2011-12-25	29

```
1  SELECT count(*) FROM mydb.tw_wrk_ics where (fecha_ingreso >= '2011-12-25' and
2  fecha_ingreso < '2012-01-01')
```

ult Set Filter: Export: Wrap Cell Content:

count(*)
29

13.3 Pruebas de informes

Verificaremos que la información obtenida en los informes es correcta.
Para ello, realizaremos unas consultas de la base de datos para verificar que la información se corresponde con la realidad.

Diagnósticos utilizados por fecha y servicio.
Verificaremos contra las tablas TW_WRK_ICS, TW_WRK_DIAGNOSTICO que los datos proporcionados son correctos.
Informe:

junio 05, 2014 @ 05:12

Diagnósticos utilizados por fecha y servicio.

Fecha Ingreso : Dec 11, 2011

Descripción del servicio	Descrición del diagnóstico
PEDIATRIA	Panencefalitis esclerosant subaguda.Enc.per cos d'inclusio Dawson,leucoenc.esclerosant Van Bogaert

Fecha Ingreso : Dec 27, 2011

Descripción del servicio	Descrición del diagnóstico
MEDICINA D'URGENCIES	Insuficiencia cardiaca esquerra. Amb cardiopatia NOS o insuf. cardiaca:edema agut pulmo,asma cardiaca
PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)

Fecha Ingreso : Dec 28, 2011

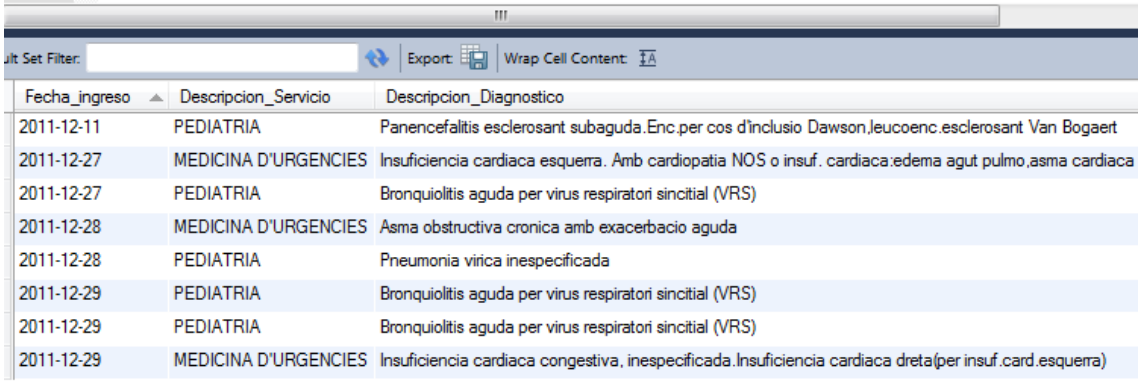
Descripción del servicio	Descrición del diagnóstico
MEDICINA D'URGENCIES	Asma obstructiva cronica amb exacerbacio aguda
PEDIATRIA	Pneumonia virica inespecificada

Fecha Ingreso : Dec 29, 2011

Descripción del servicio	Descrición del diagnóstico
PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)
PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)
MEDICINA D'URGENCIES	Insuficiencia cardiaca congestiva, inespecificada.Insuficiencia cardiaca dreta(per insuf.card.esquerra)

Datos:

```
1  SELECT a.Fecha_ingreso, b.Descripcion_Servicio, c.Descripcion_Diagnostico
2  FROM mydb.tw_wrk_ics a , mydb.tw_wrk_servicio b, mydb.tw_wrk_diagnostico c
3  where a.fecha_ingreso <= '2011-12-29'
4  and a.Id_Servicio_Entrada = b.Id_Servicio
5  and a.Id_Diagnostico = c.Id_Diagnostico
```



ult Set Filter: | Export: | Wrap Cell Content:

Fecha_ingreso	Descripcion_Servicio	Descripcion_Diagnostico
2011-12-11	PEDIATRIA	Panencefalitis esclerosant subaguda.Enc.per cos d'inclusio Dawson,leucoenc.esclerosant Van Bogaert
2011-12-27	MEDICINA D'URGENCIES	Insuficiencia cardiaca esquerra. Amb cardiopatia NOS o insuf. cardiaca:edema agut pulmo,asma cardiaca
2011-12-27	PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)
2011-12-28	MEDICINA D'URGENCIES	Asma obstructiva cronica amb exacerbacio aguda
2011-12-28	PEDIATRIA	Pneumonia virica inespecificada
2011-12-29	PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)
2011-12-29	PEDIATRIA	Bronquiolitis aguda per virus respiratori sincitial (VRS)
2011-12-29	MEDICINA D'URGENCIES	Insuficiencia cardiaca congestiva, inespecificada.Insuficiencia cardiaca dreta(per insuf.card.esquerra)

Frecuencia de atenciones por servicio y fecha.
Verificaremos contra la tabla TC_WRK_ICS que los datos proporcionados son correctos.

junio 05, 2014 @ 05:32

Frecuencia de atenciones por servicio y fecha

Fecha Intervalo : Dec 11, 2011

Descripción del Servicio	Frecuencia
PEDIATRIA	1

Fecha Intervalo : Dec 25, 2011

Descripción del Servicio	Frecuencia
PEDIATRIA	9
MEDICINA D'URGENCIES	20

Fecha Intervalo : Jan 1, 2012

Descripción del Servicio	Frecuencia
ANGIOLOGIA I CIRURGIA VASCULAR	1
CIRURGIA GENERAL I DIGESTIVA	9
CIRURGIA ORT I TRAUMATOLOGIA	24
GINECOLOGIA	116
NEUROCIRURGIA	5
NEUROLOGIA	23
OBSTETRICIA	4
PEDIATRIA	488
MEDICINA D'URGENCIES	563
UROLOGIA	5

Período que comienza el 11-12-2011:

```
1  SELECT b.Descripcion_Servicio , count(*)
2  FROM mydb.tw_wrk_ics a , mydb.tw_wrk_servicio b
3  where a.fecha_ingreso >= '2011-12-11'
4  and  a.fecha_ingreso < '2011-12-18'
5  and a.Id_Servicio_Entrada = b.Id_Servicio
6  group by b.Descripcion_Servicio
7
```

lt Set Filter: Export: Wrap Cell Content

Descripcion_Servicio	count(*)
PEDIATRIA	1

Período que comienza el 25-12-2011:

```
1  SELECT b.Descripcion_Servicio , count(*)
2  FROM mydb.tw_wrk_ics a , mydb.tw_wrk_servicio b
3  where a.fecha_ingreso >= '2011-12-25'
4  and  a.fecha_ingreso < '2012-01-01'
5  and a.Id_Servicio_Entrada = b.Id_Servicio
6  group by b.Descripcion_Servicio
7
```

ult Set Filter: Export: Wrap Cell Content:

Descripcion_Servicio	count(*)
MEDICINA D'URGENCIES	20
PEDIATRIA	9

Período que comienza el 01-01-2012:

```
1  SELECT b.Descripcion_Servicio , count(*)
2  FROM mydb.tw_wrk_ics a , mydb.tw_wrk_servicio b
3  where a.fecha_ingreso >= '2012-01-01'
4  and  a.fecha_ingreso < '2012-01-08'
5  and a.Id_Servicio_Entrada = b.Id_Servicio
6  group by b.Descripcion_Servicio
7
```

sult Set Filter: Export: Wrap Cell Conter

Descripcion_Servicio	count(*)
ANGIOLOGIA I CIRURGIA VASCULAR	1
CIRURGIA GENERAL I DIGESTIVA	9
CIRURGIA ORT I TRAUMATOLOGIA	24
GINECOLOGIA	116
MEDICINA D'URGENCIES	563
NEUROCIRURGIA	5
NEUROLOGIA	23
OBSTETRICIA	4
PEDIATRIA	488
UROLOGIA	5

Listado de tuits por búsqueda e idioma.
Revisaremos contra la tabla TC_ORI_TWITTER que se muestra la información correspondiente por búsqueda e idioma.

junio 05, 2014 @ 05:38

Listado de tuits por búsqueda e idioma

Palabras de búsqueda: gripe

Idioma tuit: ''
''

fecha	tuit
Jan 25, 2013	Por lo visto todos luego estamos con gripe y con la garganta hecho mierda :O
Jan 26, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Jan 30, 2013	Gripe o qu
Jan 30, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Jan 31, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Feb 1, 2013	El ventilador en 1 me da calor, pero en 2 ya me da fr?o. Eso quiere decir q sigo con fiebre o q soy hinchapelotas, nom?s? #duda #gripe
Feb 1, 2013	No s? si es peor esta gripe o verla a usted JAJAJA.
Feb 1, 2013	Ay gripe HDP me jodiste mi fin de semana fiesta!!!:O
Feb 3, 2013	el t? que estoy tomando es un asco, pero quiero que se me pase la gripe o lo que mierda tenga
Feb 5, 2013	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)
Feb 19, 2013	Que te mejore la gripe un d?a y al siguiente estar peor que hace dos &lt
Feb 19, 2013	@JoseAgd14
Feb 21, 2013	@leegchannie bien con mucho color de cabeza y creo que me va a dar gripe... todav?a no empese con exames pero la otra semana seguro.ji.
Apr 2, 2013	Se acerca el finde y yo con gripe...?

Datos correspondientes de TC_ORI_TWITTER:

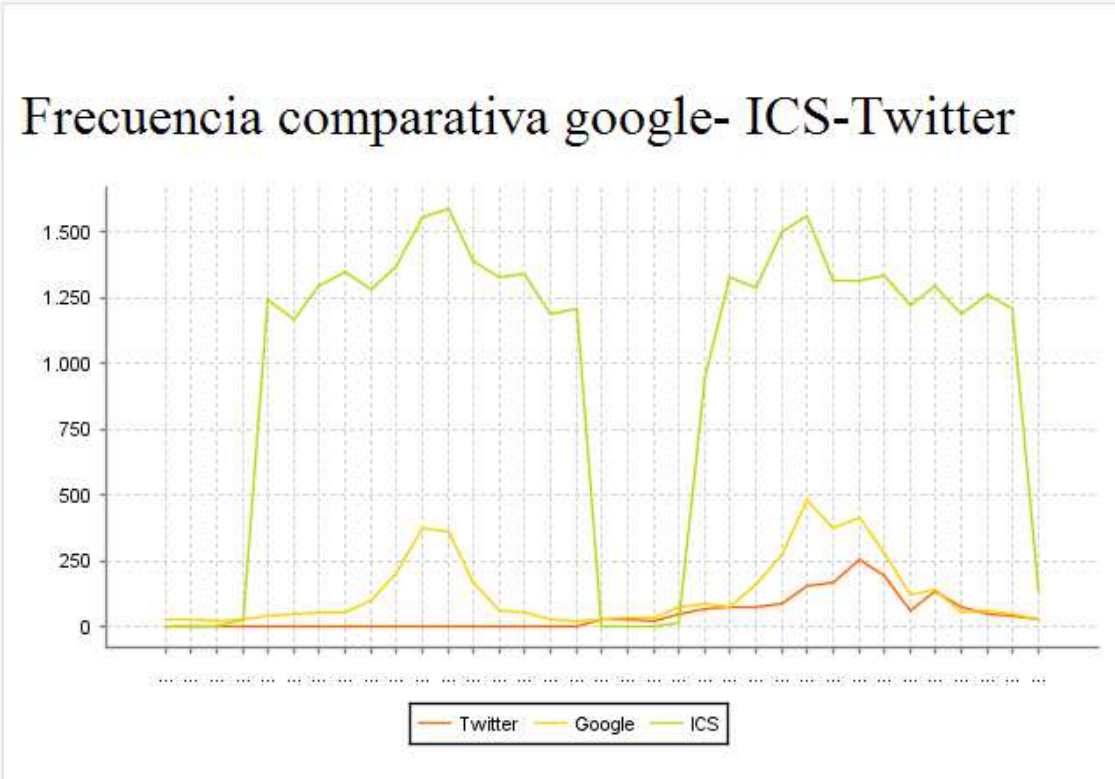
```
1  SELECT * FROM mydb.tc_ori_twitter
2  where busqueda = 'gripe' and idioma = '\r'
3
```

Set Filter: Export: Wrap Cell Content:

Fecha	Tuit	Busqueda	Idioma
2013-01-25	Por lo visto todos luego estamos con gripe y con la garganta hecho mierda :O	gripe	
2013-01-26	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)	gripe	
2013-01-30	Gripe o qu	gripe	
2013-01-30	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)	gripe	
2013-01-31	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)	gripe	
2013-02-01	El ventilador en 1 me da calor, pero en 2 ya me da fr?o. Eso quiere decir q sigo con fiebre o q soy hinchapelotas, nom?s? #duda #gripe	gripe	
2013-02-01	No s? si es peor esta gripe o verla a usted JAJAJA.	gripe	
2013-02-01	Ay gripe HDP me jodiste mi fin de semana fiesta!!!:O	gripe	
2013-02-03	el t? que estoy tomando es un asco, pero quiero que se me pase la gripe o lo que mierda tenga	gripe	
2013-02-05	Los q' viajan a China deben vacunarse contra la gripe aviar, s? o s? (Dr. Stamboulian en La Otra Agenda)	gripe	
2013-02-19	Que te mejore la gripe un d?a y al siguiente estar peor que hace dos &lt	gripe	
2013-02-19	@JoseAgd14	gripe	
2013-02-21	@leegchannie bien con mucho color de cabeza y creo que me va a dar gripe... todav?a no empese con exames pero la otra semana seguro.ji.	gripe	
2013-04-02	Se acerca el finde y yo con gripe...?	gripe	

Frecuencia comparativa Google-ICS-Twitter.

Para este informe debemos comparar las tablas TC_WRK_ICS, TC_WAR_GOOGLE y TC_WRK_TWITTER. Debemos verificar que las frecuencias de todas las tablas coincidan con la presentada en el gráfico.



Es complicado revisar en este tipo de gráfico los datos mostrados, máxime cuando el número de columnas mostradas es alto (35).

En lugar de ello, vamos a revisar con atención el origen de datos, verificando que los datos recuperados son exactamente lo que queremos mostrar.

Verificamos que la selección sobre las tablas origen es la esperada, y que si ejecutamos el SQL contra la tabla obtenemos los datos esperados.

```
1  select 'ICS' as origen, fecha_intervalo, frecuencia from mydb.tw_war_ICS
2  where (fecha_intervalo > '2011-12-01' and fecha_intervalo < '2012-04-01' ) or
3  (fecha_intervalo > '2012-12-01' and fecha_intervalo < '2013-04-01' )
4  union
5  select 'Twitter' as origen, a.fecha_intervalo, a.frecuencia from mydb.tw_war_twitter a
6  where (fecha_intervalo > '2011-12-01' and fecha_intervalo < '2012-04-01' ) or
7  (fecha_intervalo > '2012-12-01' and fecha_intervalo < '2013-04-01' )
8  union
9  select 'Google' as origen, b.fecha_intervalo, b.frecuencia from mydb.tw_war_google b
10 where (fecha_intervalo > '2011-12-01' and fecha_intervalo < '2012-04-01' ) or
11 (fecha_intervalo > '2012-12-01' and fecha_intervalo < '2013-04-01' )
12 order by Fecha_intervalo ;
13
```

ult Set Filter: Export: Wrap Cell Content:

origen	fecha_intervalo	frecuencia
Twitter	2012-12-30	71
Google	2012-12-30	91
ICS	2012-12-30	949
Twitter	2013-01-06	78
Google	2013-01-06	75
ICS	2013-01-06	1326
Twitter	2013-01-13	73
Google	2013-01-13	164
ICS	2013-01-13	1290
ICS	2013-01-20	1502
Twitter	2013-01-20	90
Google	2013-01-20	277
ICS	2013-01-27	1558
Twitter	2013-01-27	156
Google	2013-01-27	482

13.4 Pruebas del cuadro de mandos.

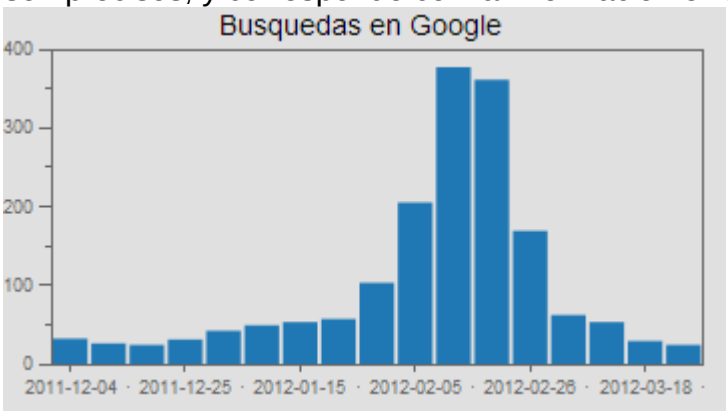
En el cuadro de mandos poseemos las siguientes informaciones:
-Frecuencias de búsqueda en google por temporada.
-Frecuencias de utilización del ICS por temporada.
-Frecuencias mensuales de búsqueda en google.
-Datos semanales: actual, anterior, media.
-Cálculo de los KPI.
Las pruebas consistirán en verificar que todos estos parámetros son correctos.

El cuadro de mandos a revisar es el siguiente:

Vamos a verificar por partes la información que se nos muestra en dicho cuadro de mandos, para comprobar la exactitud de los datos mostrados.

Frecuencias de búsqueda en google por temporada

Revisamos que los datos mostrados sobre la frecuencia de consulta de google son precisos, y corresponde con la información existente en la base de datos.



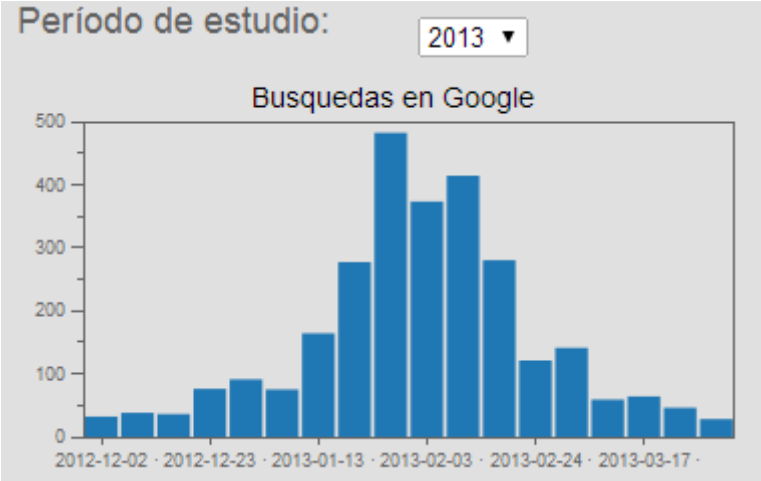

```
1  select b.fecha_intervalo, b.frecuencia from mydb.tw_war_google b
2  where (fecha_intervalo > '2011-12-01' and fecha_intervalo < '2012-04-01' )
3  order by Fecha_intervalo ;
4
```

Set Filter: Export: Wrap Cell Content:

fecha_intervalo	frecuencia
2011-12-04	32
2011-12-11	26
2011-12-18	24
2011-12-25	31
2012-01-01	42
2012-01-08	49
2012-01-15	53
2012-01-22	57
2012-01-29	103
2012-02-05	205
2012-02-12	377
2012-02-19	361
2012-02-26	169
2012-03-04	62
2012-03-11	53
2012-03-18	29
2012-03-25	24

Podemos comprobar que la información existente en TW_WAR_GOOGLE coincide con la mostrada en el gráfico para dicho período.

Además, verificamos que el gráfico responde a la selección de temporada correctamente, mostrando los datos adecuados:



Frecuencias atenciones en el ICS por temporada

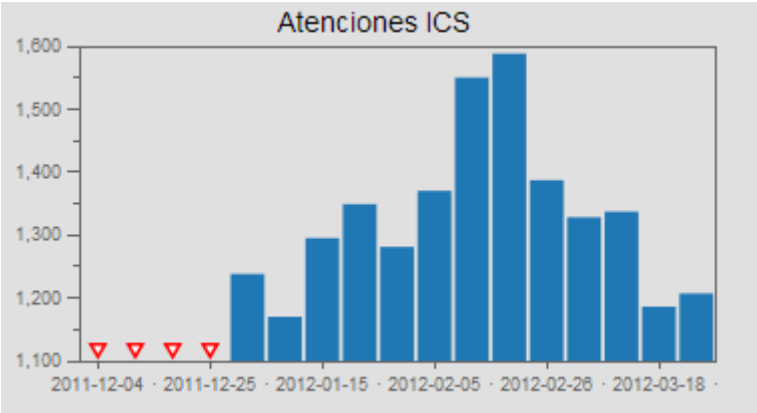
En el siguiente gráfico tenemos una muestra de la variación de las atenciones en el ICS según la temporada elegida.

Podemos apreciar que se trata de un gráfico sesgado, con un número mínimo de 1100.

Dicho sesgo se ha elegido para destacar aún más las atenciones superiores a la media, ya que de otra manera existe muy poca diferencia visual entre una frecuencia que obligue a tomar medidas extraordinarias (sobre las 1500 atenciones/semana) de una atención normal de otro período (rondando sobre las 1100 atenciones/semana).

De esta manera es mucho más sencillo apreciar los picos en la atención prestada.

Los datos inferiores a 1100 atenciones se muestran con un pequeño icono rojo, indicando que el dato es inferior al mínimo.



Comprobamos los datos existentes en la base de datos:

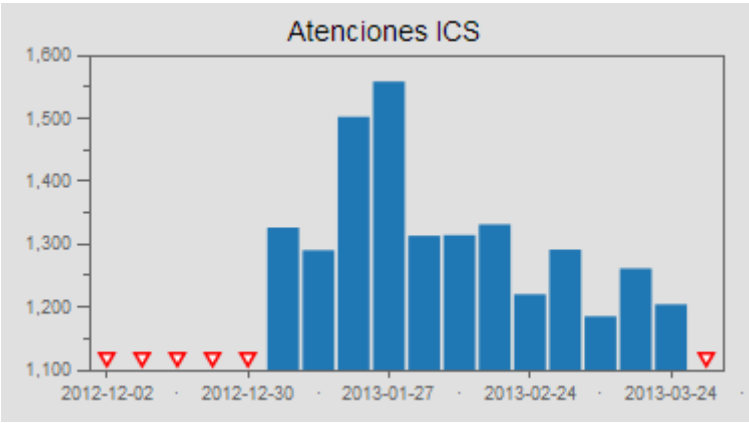
```
1  select fecha_intervalo, frecuencia from mydb.tw_war_ICS
2  where (fecha_intervalo > '2011-12-01' and fecha_intervalo < '2012-04-01' )
3  order by Fecha_intervalo ;
```

ult Set Filter: Export: Wrap Cell Content:

fecha_intervalo	frecuencia
2011-12-04	0
2011-12-11	1
2011-12-18	0
2011-12-25	29
2012-01-01	1238
2012-01-08	1170
2012-01-15	1295
2012-01-22	1349
2012-01-29	1281
2012-02-05	1370
2012-02-12	1550
2012-02-19	1588
2012-02-26	1387
2012-03-04	1328
2012-03-11	1337
2012-03-18	1186
2012-03-25	1207

Verificamos que los gráficos son correctos, ya que la información mostrada se corresponde con la existente en la base de datos.

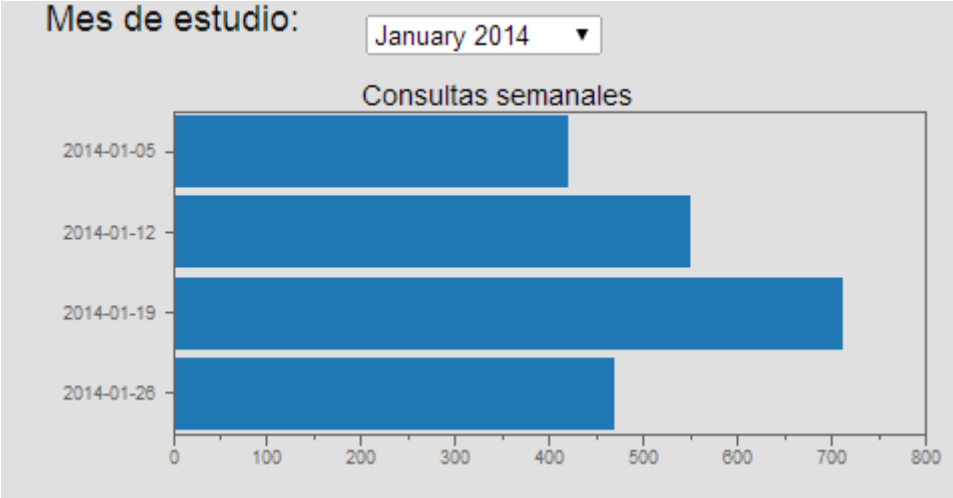
Además, verificamos que el gráfico responde a la selección de temporada correctamente, mostrando los datos adecuados:

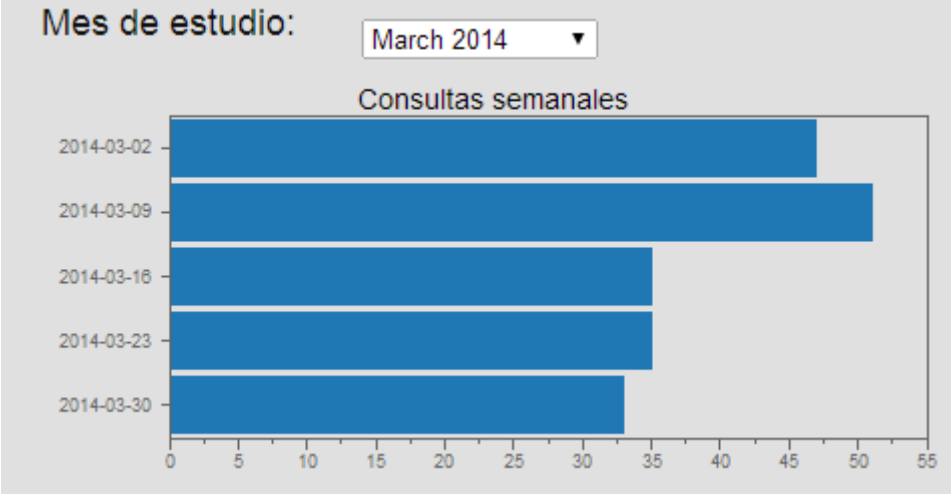


Detalle del mes en curso

Sin duda, una de las herramientas fundamentales del proyecto, puesto que nos muestra información sobre la temporada actual, permitiendo elegir el mes de estudio.

Verificamos que el gráfico responde correctamente al selector, mostrándonos información sobre el mes seleccionado:





Revisamos que la información mostrada corresponde con la existente en la base de datos:

```
1 •  SELECT * FROM mydb.tw_war_google
2    where fecha_intervalo >= '2014-03-01';
```

Result Set Filter: Export:

Fecha_Intervalo	Frecuencia
2014-03-02	47
2014-03-09	51
2014-03-16	35
2014-03-23	35
2014-03-30	33

Información detallada sobre un período seleccionado:
Esta información aparece al clicar sobre una barra del gráfico anterior, y tan sólo en esa ocasión.
Revisamos que la información es correcta, y que los KPI se cumplen.
Incluiremos la información de la frecuencia de dos semanas antes de la fecha, ya que, aunque no aparece por pantalla, se toma en cuenta para los cálculos.

Seguidamente vamos a revisar la información generada en varios períodos, con diferentes resultados en cada uno de ellos.

- Semana del 2013-12-01:

Revisamos que los datos coincidan con los existentes en la base de datos.

```
SELECT 'actual' as tipo, frecuencia FROM mydb.tw_war_google
  where fecha_intervalo = '2013-12-01'
union
SELECT 'actual -1' as tipo, frecuencia FROM mydb.tw_war_google
  where fecha_intervalo = subdate('2013-12-01', INTERVAL 7 DAY)
union
SELECT 'actual -2' as tipo, frecuencia FROM mydb.tw_war_google
  where fecha_intervalo = subdate('2013-12-01', INTERVAL 14 DAY)
union
SELECT 'media', round(avg(frecuencia)) as frecuencia FROM mydb.tw_war_google
 where fecha_intervalo >= '2013-12-01' and fecha_intervalo <= '2014-03-31'
```

ult Set Filter: Export: Wrap Cell Content:

tipo	frecuencia
actual	43
actual -1	39
actual -2	35
media	197

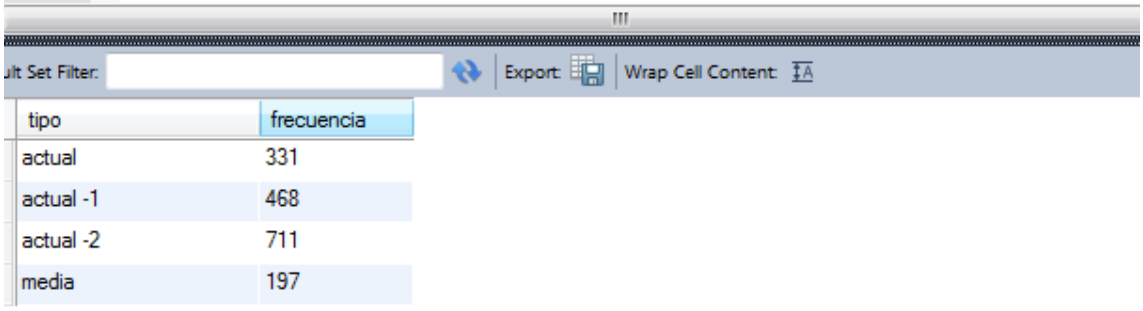
Comprobamos el KPI según los valores del apartado Creación de un cuadro de mando. :

La frecuencia actual (43) es **inferior a la media (197)**, por lo que el riesgo se califica como **bajo**

- Semana del 2014-02-02:

Revisamos que los datos coincidan con los existentes en la base de datos.

```
1  SELECT 'actual' as tipo, frecuencia FROM mydb.tw_war_google
2    where fecha_intervalo = '2014-02-02'
3  union
4  SELECT 'actual -1' as tipo, frecuencia FROM mydb.tw_war_google
5    where fecha_intervalo = subdate('2014-02-02', INTERVAL 7 DAY)
6  union
7  SELECT 'actual -2' as tipo, frecuencia FROM mydb.tw_war_google
8    where fecha_intervalo = subdate('2014-02-02', INTERVAL 14 DAY)
9  union
10 SELECT 'media', round(avg(frecuencia)) as frecuencia FROM mydb.tw_war_google
11  where fecha_intervalo >= '2013-12-01' and fecha_intervalo <= '2014-03-31'
12
```



tipo	frecuencia
actual	331
actual -1	468
actual -2	711
media	197

Comprobamos el KPI según los valores del apartado Creación de un cuadro de mando. :

La frecuencia actual (331) es **superior a la media (197)** , **pero inferior a 1.75 veces la frecuencia de la semana anterior (702)** por lo que el riesgo se califica como **medio**

- Semana del 2013-12-29:

Revisamos que los datos coincidan con los existentes en la base de datos.

```
1  SELECT 'actual' as tipo, frecuencia FROM mydb.tw_war_google
2    where fecha_intervalo = '2013-12-29'
3  union
4  SELECT 'actual -1' as tipo, frecuencia FROM mydb.tw_war_google
5    where fecha_intervalo = subdate('2013-12-29', INTERVAL 7 DAY)
6  union
7  SELECT 'actual -2' as tipo, frecuencia FROM mydb.tw_war_google
8    where fecha_intervalo = subdate('2013-12-29', INTERVAL 14 DAY)
9  union
10 SELECT 'media', round(avg(frecuencia)) as frecuencia FROM mydb.tw_war_google
11  where fecha_intervalo >= '2013-12-01' and fecha_intervalo <= '2014-03-31'
12
```

ult Set Filter: Export: Wrap Cell Content:

tipo	frecuencia
actual	309
actual -1	108
actual -2	59
media	197

Comprobamos el KPI según los valores del apartado Creación de un cuadro de mando. :

La frecuencia actual (309) es **superior a la media (197)** , **y a 1.75 veces la frecuencia de la semana anterior (189)** por lo que el riesgo se califica como **alta.**
Esta frecuencia nos indicaría el comienzo de un brote de gripe.

- Semana del 2014-01-05:

Revisamos que los datos coincidan con los existentes en la base de datos.

```
1  SELECT 'actual' as tipo, frecuencia FROM mydb.tw_war_google
2    where fecha_intervalo = '2014-01-05'
3  union
4  SELECT 'actual -1' as tipo, frecuencia FROM mydb.tw_war_google
5    where fecha_intervalo = subdate('2014-01-05', INTERVAL 7 DAY)
6  union
7  SELECT 'actual -2' as tipo, frecuencia FROM mydb.tw_war_google
8    where fecha_intervalo = subdate('2014-01-05', INTERVAL 14 DAY)
9  union
10 SELECT 'media', round(avg(frecuencia)) as frecuencia FROM mydb.tw_war_google
11  where fecha_intervalo >= '2013-12-01' and fecha_intervalo <= '2014-03-31'
12
```

ult Set Filter: Export: Wrap Cell Content:

tipo	frecuencia
actual	420
actual -1	309
actual -2	108
media	197

Comprobamos el KPI según los valores del apartado Creación de un cuadro de mando. :

La frecuencia actual (420) y la de la semana anterior (309) son **superiores a la media (197)** . Además, ambas semanas son superiores **a 1.75 veces la frecuencia de hace dos semanas (189)** por lo que el riesgo se califica como **muy alto.**
Esta frecuencia nos indicaría la segunda semana de un brote de gripe.

De hecho, hemos marcado el período anterior como el inicio de un brote, con lo cual podemos validar la coherencia del proceso.

14 Anexo 2: Pruebas de usabilidad.

Hemos realizado las pruebas de usabilidad reseñadas en el apartado Pruebas de usabilidad.
Se realiza dicha prueba con un sujeto de pruebas que podría modelizar un usuario típico del sistema.

14.1Datos pre-test

Como datos pre-test, apuntamos que el sujeto de pruebas es un varón de 17 años con buenos conocimientos de informática, aunque sin experiencia concreta en sistemas BI.
Tampoco posee conocimientos específicos del funcionamiento del sistema de salud, o del seguimiento y predicción por medios informáticos de brotes de gripe.

14.2 Realización del test.

Se realiza un vídeo con la prueba. Dicho vídeo puede consultarse en:
https://dl.dropboxusercontent.com/u/4908045/Prueba%20usuario.mp4

Tras el análisis del vídeo se extraen las siguientes conclusiones:

-El número de datos presentados por pantalla son correctos. No se llega a sepultar los datos más interesantes con una amalgama de otros no relacionados, consiguiendo así el correcto protagonismo de cada uno.

-La presentación por pantalla es suficientemente clara y concisa.

-Los informes son correctos, aunque no tienen gran significado para un usuario que no conozca el funcionamiento del ICS. Quizás los usuarios finales podrían aportar una serie de modificaciones o nuevos informes que pudieran ser más adecuados a sus estilos de trabajo habituales.

-En general, la funcionalidad del cuadro de mandos se ve buena.

14.3 Cuestionario post-test y autorizaciones.

Se realiza el siguiente cuestionario post-test para clarificar las impresiones del sujeto de pruebas.

Evalúe las siguientes afirmaciones de 1-10 donde 1 significa nada de acuerdo y 10 totalmente de acuerdo.

-Los informes y cuadro de mando son sencillos de utilizar.
-He podido realizar las tareas encomendadas de manera fácil e intuitiva.
-En todo momento sabía el estado de la tarea que estaba realizando.
-En caso de error, he tenido la información suficiente para corregirlo.
-La información se me ha mostrado de manera clara y sencilla.
-Los formularios son cortos y fáciles de entender.
-La experiencia ha sido cómoda, y el proceso, rápido.

Dado que hemos grabado en vídeo la prueba, se solicita la siguiente autorización:

Autorización para la grabación en vídeo:

Yo, Cristian Hernández Hernández, con DNI 53751798Q autorizo específicamente a Sergio Hernández Gasó , NIF 53053575G a que pueda ser grabado en vídeo exclusivamente para el análisis del estudio del que me han informado.
En el caso de que quiera revocar dicho consentimiento, he de avisar por escrito a Sergio Hernández Gasó para que eliminen completamente el material obtenido de esta manera.
También acepto que la información intercambiada es propiedad de Sergio Hernández Gasó y me comprometo a no divulgarla si no poseo el expreso consentimiento del mismo.