Solución

Àlex Bartrolí Muñoz

PID_00209809

UOC
Universitat Oberta de Catalunya

Índice

Introducción

Tal y como ya anunciábamos en el enunciado del caso, la elaboración de herramientas de minería y explotación de la información supone un trabajo de meses o incluso años, con la participación de equipos multidisciplinarios que los van implementando a lo largo del tiempo, en un proceso de mejora continua.

Por tanto, el objetivo de este caso, no es crear una solución completa, sino que el estudiante utilice las distintas herramientas y metodologías para desarrollar un sistema capaz de predecir los brotes víricos de gripe con mayor antelación.

La resolución de esta actividad, "detección de brotes víricos a partir de redes sociales", consta de tres partes:

1) Recuperación de la información: donde se analizan las diferentes herramientas para recuperar la información de Twitter, se indica cómo se puede filtrar la información y qué consultas deben realizarse.

2) Creación de un cuadro de mando: donde se definirá un sistema de indicadores que permita comparar el impacto actual de la gripe en Twitter respecto al impacto en Twitter en otros periodos pasados que hayan precedido a un brote del virus de la gripe.

3) Carga y explotación de los datos: donde el objetivo final es proponer un sistema de almacenamiento de datos que permita realizar el seguimiento estratégico y operativo de los mensajes escritos en Twitter y la información disponible de los hospitales.

1. Recuperación de los mensajes sobre la gripe en Twitter

La primera parte de este proyecto analiza las diferentes herramientas para recuperar datos de Twitter, las características de estos datos, cómo podemos segmentarlos y las restricciones de cada una de las herramientas.

Para escoger una herramienta de recuperación de mensajes sobre la gripe debemos tener en cuenta que:

- Las herramientas para acceder a los datos de Twitter y las restricciones que tienen cada una de ellas.
- Los parámetros disponibles para seleccionar los datos que se han de recuperar.
- Las búsquedas que se realizarán y la cantidad de mensajes devueltos con esas búsquedas.

1.1. Herramientas de recuperación de datos

Para recuperar los datos de Twitter se pueden utilizar bibliotecas que interactúen con la API de Twitter, ayudarse de alguna herramienta web o utilizar la búsqueda avanzada de Twitter o la consola de Twitter para desarrolladores. Para seleccionar la mejor opción exploraremos los pros y los contras de cada posible solución.

Herramientas de web

Hay muchas webs que permiten recuperar y analizar datos de Twitter y estas están en continua evolución. En este enlace tenéis una lista de algunas de ellas.

La interfaz web de Twitter y su búsqueda avanzada son una buena opción para realizar una primera exploración de forma rápida. Si queremos validar si una palabra clave de búsqueda es utilizada por un número adecuado de usuarios, la mejor opción es realizar la búsqueda en la web de Twitter y explorar los resultados de forma visual e interactiva. Las opciones de la búsqueda avanzada nos permiten aplicar filtros en las consultas, como por ejemplo el idioma de los tuits o el intervalo de tiempo en que fueron escritos.

En cambio, la interfaz web solo nos sirve para realizar una primera exploración visual, ya que no permite automatizar la extracción y almacenamiento de datos, la segmentación geográfica no funciona correctamente y existen otras herramientas que nos permiten aplicar más filtros de utilidad para segmentar las consultas realizadas a Twitter.

La consola de Twitter para desarrolladores tiene la ventaja de que permite utilizar los mismos parámetros de segmentación que la API de Twitter y devuelve de forma visual los mismos resultados que obtendríamos con una consulta a la API.

Por el contrario el resultado devuelto es una estructura JSON con más de 100 parámetros que resulta muy difícil de analizar a simple vista y tampoco permite guardar los resultados de forma automática.

Una tercera opción es utilizar una herramienta comercial ya disponible que se adapte a nuestras necesidades. Hay muchas herramientas de recuperación de mensajes de Twitter y, teniendo en cuenta que las herramientas comerciales han sido ampliamente testadas, utilizar una herramienta externa puede ser una buena opción.

También debemos tener en cuenta los inconvenientes que puede tener una herramienta externa:

- No está garantizada la continuidad de la herramienta. Una cantidad muy importante de herramientas cierra sus puertas en un plazo inferior a dos años.

- Tampoco está garantizada la actualización de la herramienta para adaptarse a las novedades en los datos que nos permite recuperar Twitter.

Para finalizar exploraremos la API de Twitter. La API de Twitter cumple todas las necesidades de nuestro proyecto, ya que permite aplicar los filtros necesarios para segmentar la recuperación de datos, permite recuperar los mensajes escritos sobre la gripe de forma automatizada y realizar las consultas en intervalos cortos de tiempo.

La API de búsqueda de Twitter ofrece una serie de parámetros para filtrar los datos que recuperar, entre los que destacan:

Parámetros de búsqueda

Podéis ver todos los parámetros para personalizar las búsquedas en este enlace.

Idiomas

En este enlace se pueden encontrar los idiomas con los que se puede personalizar la búsqueda en Twitter. “en” para inglés, “es” para español, “fr” para francés, etc. En catalán no está disponible.

Tabla 1. Parámetros de búsqueda más relevantes

Nombre del parámetro	Significado
Q (obligatorio)	La cadena de texto que se busca
Geocode	Devuelve solo los tuits de los usuarios localizados en un radio *x* de las coordenadas geográficas indicadas
lang	Devuelve solo los tuits en el idioma indicado
locale	Indica el idioma utilizado en la cadena de texto que se busca
result_type	Indica el tipo de resultados que se desean recibir. El valor por defecto es *mixed* y las posibilidades son: • *recent*: los resultados más recientes • *popular*: los resultados más populares • *mixed*: tuits populares y recientes
Until	Tuits generados hasta la fecha indicada
Since	Tuits generados desde la fecha indicada

Con estos parámetros podemos filtrar los tuits por idioma, por fecha y recuperar solo aquellos mensajes ubicados en una cierta área geográfica, aunque debemos tener en cuenta las siguientes restricciones de la API.

- Devuelve un máximo de 100 mensajes por petición, y solo almacena los mensajes un máximo de 10 días[1].

(1)Twitter no indica esta restricción en la especificación de su API, pero desde que salió la última versión de su API, se pueden encontrar muchas quejas de muchos usuarios al respecto y todos coinciden en que no pueden recuperar mensajes de hace 10 o más días.

- El número de peticiones está limitado. Los límites varían en función de la opción de la consulta utilizada y la opción de búsqueda de tuits tiene un límite de 540 peticiones cada 15 minutos, es decir, podemos realizar, de media, una búsqueda cada 2 segundos.

Límites de consulta

Los límites de las consultas que hay que realizar por intervalos de tiempo están disponibles en este enlace.

Una vez vistas las posibilidades de recuperación de información y las restricciones del API de Twitter, y teniendo en cuenta que las herramientas realizadas por terceros utilizarán la API de Twitter para obtener la información, hemos considerado que la API de Twitter es la mejor opción para recuperar los mensajes escritos por los usuarios sobre la gripe.

1.2. Parámetros disponibles para segmentar la recuperación de datos

Una vez escogida la API de Twitter y conociendo sus filtros de búsqueda, definiremos las posibles búsquedas que realizar para recuperar los mensajes de los usuarios. Para ello tendremos en cuenta:

- El/los idioma/s en que se realizarán las búsquedas.
- La/s zona/s geográficas donde fueron escritos los mensajes.
- Las cadenas de texto que se buscan.
- La periodicidad en la que recuperaremos la información.

1.2.1. Idiomas de las búsquedas

Lo primero que haremos será plantearnos con que idiomas vamos a trabajar. Dado que la herramienta será utilizada en el sistema sanitario catalán, los idiomas seleccionados adecuados serían catalán, castellano e inglés, pero como el Twitter no permite filtrar los mensajes en catalán, trabajaremos con castellano e inglés.

1.2.2. Localización geográfica de los mensajes

El siguiente paso será limitar el área geográfica de la que se recuperará la información. Para ello escogeremos un punto céntrico y aplicaremos un radio de cobertura. Teniendo en cuenta que la zona más poblada del ámbito de actuación del sistema sanitario catalán es Barcelona y su metrópoli, y que, en

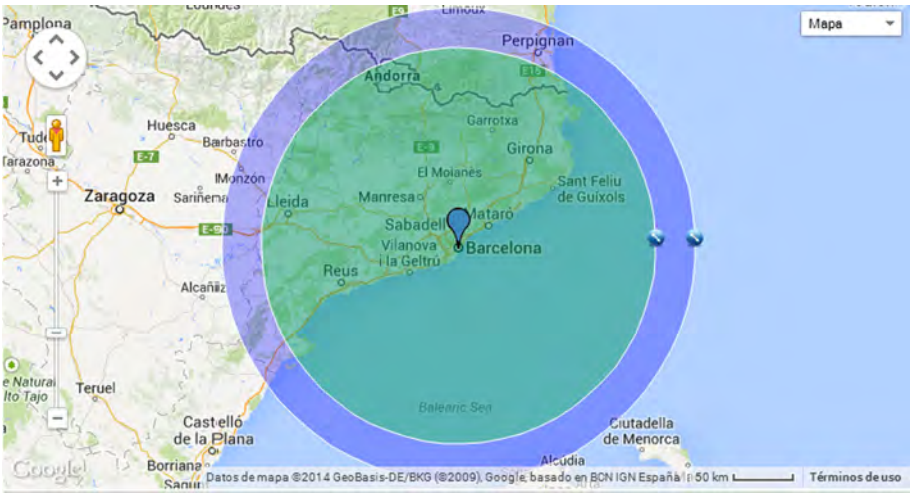
principio, debería ser la zona desde la que se escribirán más mensajes sobre la gripe en Twitter, escogeremos este lugar como punto de partida y ampliaremos la búsqueda en un radio de 180 km alrededor, para cubrir todo el ámbito de actuación del sistema sanitario catalán.

Filtro de zona geográfica

En la API de Twitter, el filtro de zona geográfica se realiza mediante las coordenadas geográficas de un punto y el radio de cobertura expresado en kilómetros o en millas. La latitud y longitud de Barcelona son: 41.3825 y 2.176944.

También podríamos haber escogido varios lugares con un radio de actuación menor desde los que recuperar la información de Twitter. Por simplicidad se ha decidido escoger un solo punto de origen. Hay que tener presente que las restricciones de la API de Twitter pueden variar y se podría limitar el radio de recuperación de los mensajes. En tal caso se deberán escoger el mínimo conjunto de coordenadas geográficas y un radio de actuación para cada una de ellas, de manera que permitan cubrir todo el ámbito de actuación del sistema sanitario catalán. Hay que tener en cuenta que habrá algún solapamiento entre las regiones escogidas y, por tanto, se deberán descartar los mensajes repetidos.

Figura 1. Zona de cobertura de los mensajes recuperados de Twitter con centro en Barcelona y un radio de 180 km y 200 km



1.2.3. Cadenas de búsqueda

El siguiente paso consistirá en elaborar una lista de los términos utilizados por los usuarios para hablar de la gripe.

Para ello, elaboraremos una lista de términos iniciales y verificaremos con la interfaz web o la consola de desarrolladores si los términos son utilizados por los usuarios de forma frecuente para hablar sobre la gripe.

En castellano, una posible lista inicial estaría formada por "gripe, gripazo y engripado". La API de Twitter no distingue entre mayúsculas y minúsculas, pero devolverá un conjunto de mensajes diferentes ante búsquedas con términos como "gripe" y "gripee". Por tanto, a nuestra lista de términos, tendremos que añadir términos como "gripee", "gripeee" "gripazoo", etc.

En inglés, realizaremos la búsqueda con el término "flu" y descartaremos otras palabras utilizadas para referirse a la gripe como *ache*, *illness*, *malady* por ser demasiado genéricas y no referirse exclusivamente al virus de la gripe.

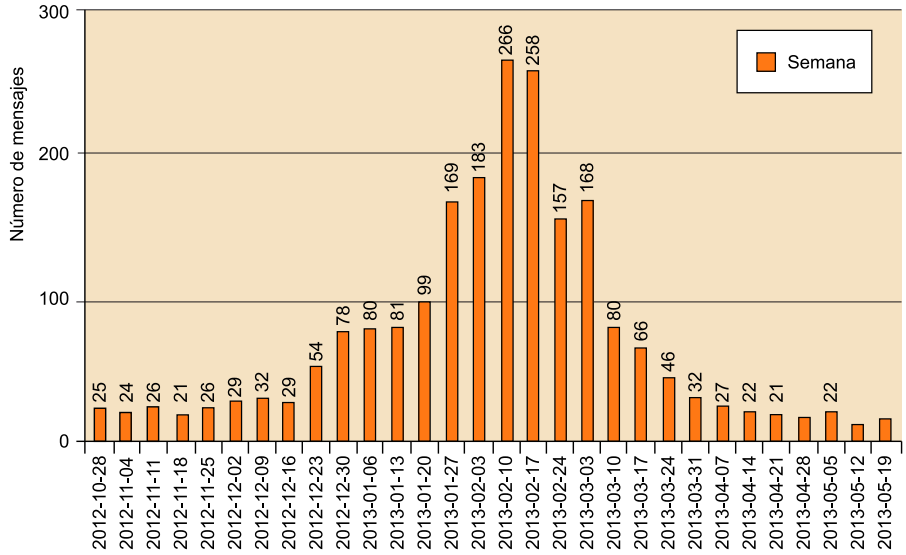
Una vez tengamos una lista de términos, deberemos verificar utilizando la búsqueda avanzada de Twitter que los términos seleccionados realmente son los que utilizan los usuarios para hablar sobre la gripe, que hay un número de mensajes suficientes para que haya capacidad predictiva y que los términos no son utilizados para hablar de otras temáticas.

Si buscamos el término "flu" en inglés, veremos que hay un elevado número de mensajes al día que utilizan este término. En cambio, si delimitamos la región de búsqueda al área de influencia del Institut Català de la Salut, el número de mensajes escritos en Twitter en inglés utilizando el término "flu" será demasiado reducido para utilizarse para prever un brote vírico de gripe.

Para aquellos términos que en la exploración visual parezca que pueden ayudar a predecir un futuro brote de gripe, deberemos recuperar un histórico de datos de años anteriores y deberemos almacenar los datos actuales para un futuro análisis. Para este caso práctico disponemos de un conjunto de mensajes de Twitter que utilizan la palabra "gripe" y el recuento semanal del número de mensajes escritos por los usuarios en diferentes fechas que nos permitirán analizar si el término "gripe" es válido para extraer la información necesaria para nuestra herramienta.

En la figura 2 observamos que el número de mensajes por semana sigue una forma de campana que puede ayudar a predecir cuándo se producirá un nuevo brote viral. En el gráfico se puede observar que antes de llegar al punto máximo se producen varios incrementos en los que se duplica el número de mensajes de una semana a la siguiente, por lo que todo parece indicar que el término "gripe" es un buen indicador para predecir un brote viral.

Figura 2. Tuits por semana para el término gripe



Fuente: Highcharts.com.

1.3. Periodicidad de recuperación de mensajes

El último aspecto que tener en cuenta es la periodicidad con la que se recuperarán los mensajes.

Si la frecuencia de recuperación es muy pequeña es posible que no recuperemos todos los mensajes escritos por los usuarios. En cambio, si recuperamos los mensajes con demasiada frecuencia, la mayoría se tendrá que descartar por estar repetidos.

Por tanto se deberán recuperar los mensajes de Twitter de forma que obtengamos todos los mensajes y realizando el número mínimo de consultas.

En el caso de que tengamos que recuperar los datos de Twitter por primera vez, escogeremos una frecuencia elevada para no perder ningún mensaje escrito por los usuarios. Una vez tengamos un histórico de datos suficiente, recuperaremos el día en el que los usuarios han escrito más mensajes sobre la gripe y dividiremos este número por el máximo de mensajes que se pueden recuperar en cada consulta.

$$\text{Número de consultas por día} = \frac{\text{Máximo de mensajes en un día}}{\text{Máximo de mensajes por consulta}}$$

Una ratio superior a 1 indica se deberá realizar más de una consulta a la API de Twitter para poder recuperar todos los mensajes escritos por los usuarios ese día.

Tal y como hemos visto en la especificación del API de Twitter, la opción de búsqueda de tuits tiene un límite de 540 peticiones cada 15 minutos, es decir, podemos realizar de media una búsqueda cada 2 segundos, lo que supone que durante un día se puede realizar un total de 43.200 consultas al API[2]. Si la ratio anterior fuera superior a 43.200 deberíamos utilizar más de una aplicación para poder recuperar todos los datos sobre la gripe escritos por los usuarios.

(2)Las 43.200 consultas se obtienen de dividir los 86.400 segundos que tiene un día entre 2, que es la frecuencia máxima de consultas al API de Twitter.

En la sección de la localización geográfica de los mensajes hemos visto que el área de actuación del Institut Català de la Salut es reducida y en el conjunto de datos que disponemos de los mensajes sobre gripe de 2013 podemos observar que el número máximo de los mensajes escritos en un mismo día utilizando el término "gripe" es de 45, por lo que con una consulta al día a la API de Twitter será suficiente para recuperar todos los mensajes escritos por los usuarios.

No obstante, debido al crecimiento de usuarios y de mensajes escritos que experimenta Twitter, en un futuro se deberá tener en cuenta que el número máximo de las peticiones diarias a la API puede variar.

2. Consulta resultante para la recuperación de datos

Para finalizar definimos cómo quedaría la consulta y cómo recuperaremos la información del API de Twitter:

Los parámetros que se utilizarán en la consulta serán:

Tabla 2. Parámetros para realizar la consulta

Nombre del parámetro	Valor
Q	Gripe
Geocode	41.3825,2.1769444444444,180km
lang	Es
locale	Recent
result_type	100
Until	El día de hoy en formato YYYY-mm-dd
Since	El día de hoy en formato YYYY-mm-dd

y la URL de llamada al API de Twitter resultante será:

```
https://api.twitter.com/1.1/search/tweets.json?
q=gripe&geocode=41.3825%2C2.1769444444444%2C180km&lang=es&
result_type=recent&count=100&until=2014-04-12&since=2014-04-12
```

Para realizar las llamadas al API utilizaremos una de las bibliotecas propuestas por Twitter para interactuar con su API, ya que la consola de desarrolladores es una herramienta de consulta visual pero no permite almacenar los datos ni realizar consultas a la API de forma automática. Las bibliotecas están disponibles en varios lenguajes de programación y las llamadas a la API de Twitter se programarán para que al finalizar el día se recuperen todos los mensajes escritos sobre la gripe en el día anterior.

3. Definición del cuadro de mando

El desarrollo de un cuadro de mando en un contexto real comporta un amplio conocimiento del entorno y un trabajo exhaustivo y multidisciplinario, en un ciclo de mejora continua. Al mismo tiempo, la consecución de un cuadro de mando es uno de los diversos resultados de una determinada estrategia (o proyecto de *outputs*), así como de su propio desarrollo y evolución en el tiempo.

Para la elaboración de un cuadro de mando antes de nada definiremos los objetivos estratégicos del ICS para la detección precoz de brotes gripales a partir de los mensajes recuperados de Twitter, el mapa estratégico con las perspectivas del cuadro de mando integral y los objetivos estratégicos. Posteriormente se analizarán posibles indicadores útiles para la toma de decisiones y se finalizará con una propuesta de cuadro de mando en forma gráfica que posibilite una rápida interpretación de los datos para la toma de decisiones.

A la hora de elaborar los indicadores para la detección de brotes de gripe a partir de los datos de Twitter se deberá tener en cuenta la tipología de los datos. Muchos de los datos utilizados en la elaboración de cuadros de mando se obtienen a partir de la interacción de la empresa con su entorno, que acumula información que puede ser fácilmente interpretable o utilizable en un cuadro de mando (número de altas, de ingresos, de defunciones...). En cambio, los tuits recuperados de Twitter se deberán analizar para poder extraer de ellos información interpretable para la toma de decisiones, ya sea de su contenido o del recuento del número de tuits escritos en intervalos de tiempo. Además, para verificar que los indicadores propuestos tienen capacidad predictiva se deberá cruzar la información obtenida a partir de los datos de Twitter con los datos disponibles sobre casos hospitalarios.

Por tanto, el objetivo del caso no es tanto desarrollar una solución completa, sino más bien hacer que el estudiante utilice las distintas herramientas y metodologías que tiene a su disposición para el desarrollo de un cuadro de mando y que sea capaz de cruzar diferentes fuentes de información para elaborar un cuadro de indicadores, que permita medir, evaluar, y, posteriormente, optimizar los procesos de negocio de acuerdo con las necesidades que vayan surgiendo en cada momento. Hay que tener presente que para la elaboración de este caso práctico disponemos de un conjunto de datos del año 2013 sobre hospitalizaciones, urgencias y tuits escritos por los usuarios sobre la gripe. Por tanto los indicadores que creemos para la detección de brotes de gripe deberán ser validados en años posteriores a partir de los nuevos datos disponibles.

Así, afrontando todos los problemas y dudas que se encontrará en el proceso de desarrollo de un caso real, el estudiante habrá podido:

- Comprobar por sí mismo la dificultad de elaborar un modelo de estas características, y todas las dudas que se generan a lo largo del proceso.

- Tomar conciencia de los múltiples factores que intervienen en una tarea de estas características.

- Identificar las bases para desarrollar un cuadro de mando de forma gradual hasta conseguir un modelo suficientemente satisfactorio y completo.

- Profundizar en todos aquellos aspectos del método que son especialmente complejos y que requieren de una amplia experiencia.

Esto sin duda le servirá para poder afrontar el desarrollo de un caso real con mayor solvencia, y a la vez, ser consciente de las dificultades y los pasos que deberá seguir.

3.1. Objetivos estratégicos en la detección de brotes víricos

Los objetivos estratégicos se centrarán en una detección más precoz de la llegada de un brote viral, una mejor gestión de los recursos y una mejor comunicación a los ciudadanos. Posibles objetivos estratégicos al respecto podrían ser:

- Predecir la aparición de nuevos brotes gripales con mayor antelación.

- Realizar una gestión más eficiente de los recursos sanitarios disponibles, tanto de personal como de material médico.

- Ofrecer una mejor información de carácter preventivo a los ciudadanos.

- Aumentar el grado de afluencia de los pacientes al centro sanitario correspondiente en función de la gravedad de la infección, así se puede incentivar que los pacientes acudan a su centro de atención primaria en vez de al hospital y que solo sean enviados al hospital aquellos pacientes que realmente lo necesiten.

Además, los objetivos estratégicos también se pueden ver desde las 4 perspectivas propuestas por Kaplan y Norton en su método para la elaboración de un cuadro de mando:

1) Perspectiva financiera:

- Reducción global de costes reduciendo el número de personas tratadas en hospitales y aumentando el número de personas que reciben tratamiento en centros de atención primaria.

- Reducción de costes a través de una mejor gestión de los recursos hospitalarios debido al mayor tiempo disponible entre la detección de la aparición real de un brote de gripe y el momento en que se producen los primeros casos de urgencias y hospitalizaciones.

2) Perspectiva del cliente o mercado:

- Permite concienciar con mayor antelación a los ciudadanos sobre la gripe y las acciones preventivas que se deben llevar a cabo para no contraer el virus.

- Permite informar a los ciudadanos con mayor antelación sobre la llegada de un nuevo brote de gripe e indicar también que la primera visita se tiene que realizar en un centro de atención primaria y que los afectados que necesiten tratamiento hospitalario serán trasladados al centro hospitalario pertinente desde el centro de atención primaria.

3) Perspectiva interna del proceso interno (o negocio):

- Se optimizan la gestión del personal sanitario y de los recursos hospitalarios debido al mayor tiempo entre la detección de la aparición real de un brote de gripe y el momento en que se producen los primeros casos de urgencias y hospitalizaciones.

- Este tiempo extra permitirá una mejor comunicación entre los diferentes centros de le red de hospitales y centros sanitarios del ICS.

4) Perspectiva del aprendizaje, crecimiento o tecnología:

- Se potencia el uso de las ciencias de la información, formando al personal sanitario y administrativo sobre nuevas técnicas de información y de análisis de datos.

- Aprovechar las innovaciones tecnológicas para optimizar los procesos de predicción de brotes víricos.

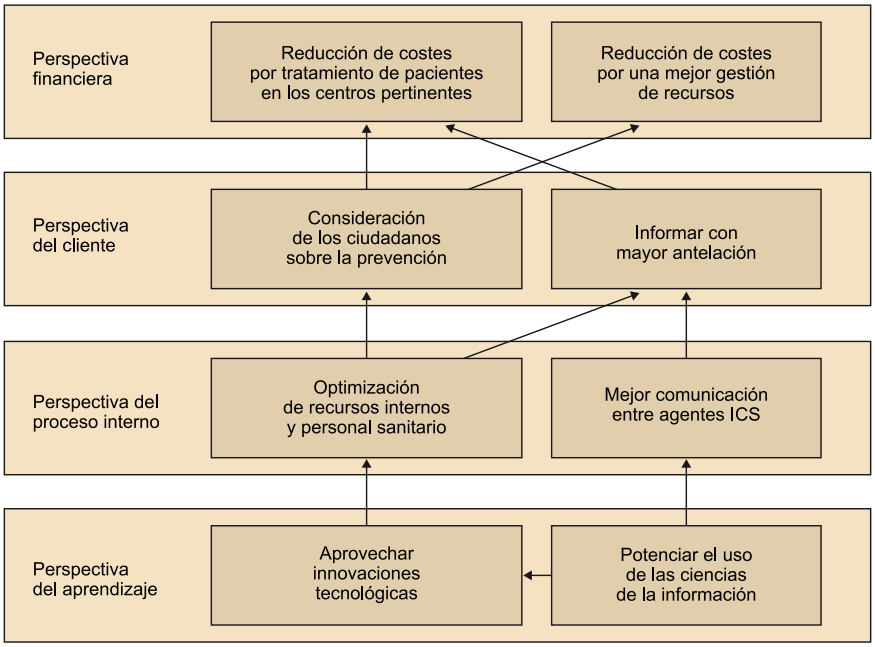
3.2. Mapa estratégico

A partir de los objetivos estratégicos descritos para el cuadro de mando integral se puede realizar un mapa estratégico. En el mapa estratégico se observan las relaciones existentes entre los diferentes objetivos estratégicos, que pueden formar parte de la misma perspectiva empresarial o de perspectivas diferentes.

A la hora de evaluar el cumplimiento de los objetivos estratégicos del mapa estratégico, se deberá empezar por la primera perspectiva, la financiera, y revisar de forma descendiente el resto de los objetivos fijados.

El mapa estratégico resultante a partir de los objetivos descritos en la visión empresarial desde las 4 perspectivas de la empresa es:

Figura 3. Mapa estratégico de las 4 perspectivas de la empresa



3.3. Fuentes de datos

Para la creación del cuadro de mandos y para la obtención y validación de indicadores que nos permitan predecir la aparición de un brote de gripe disponemos de un fichero con el conjunto de los tuits que contienen la palabra gripe entre las fechas de noviembre de 2012 y febrero de 2014. También disponemos de un fichero de datos con el histórico de hospitalizaciones y un fichero con los casos de urgencias, ambos con los casos acontecidos entre enero y marzo de los años 2012 y 2013.

3.3.1. Fichero de tuits sobre la gripe

Este fichero contiene los tuits escritos en castellano en la región de Cataluña entre noviembre de 2012 y febrero de 2014 que contienen la palabra gripe.

Cada entrada del conjunto de datos contiene los siguientes atributos:

Tabla 3. Atributos del conjunto de datos sobre la gripe

Nombre del campo	Descripción
Contenido del tuit	Contenido del mensaje de Twitter

Nombre del campo	Descripción
Tweet Link	Enlace al tuit en la web de Twitter
User	Nombre del usuario autor del tuit
Keyword	Palabra clave utilizada para realizar la búsqueda
User Link	Enlace al usuario del autor del tuit
Fecha del tuit	Fecha y hora en que fue escrito el tuit
Semana	Primer día de la semana, mes y año en que fue escrito el tuit
Idioma	Idioma del usuario que escribió el tuit
Contenido del tuit	Contenido del mensaje de Twitter

No todos los datos serán necesarios para crear los indicadores. Los datos más importantes serán el contenido de los tuits y la semana a la que pertenece cada tuit. La semana en la que son escritos los mensajes es muy importante para poder hacer un recuento de los mensajes escritos durante la semana actual y comparar este valor con el de semanas anteriores para crear indicadores que nos permitan detectar la posible llegada de un nuevo brote de gripe en un plazo corto de tiempo. El contenido del tuit también es importante porque puede haber ocasiones en que un incremento en el número de mensajes en Twitter sobre la gripe sea debido a la aparición de una noticia relacionada con la gripe o que una persona muy mediática sea noticia por algún tema relacionado con la gripe. En estos casos, una lectura rápida de los tuits nos permitirá saber si el incremento de mensajes escritos es debido a la llegada de un nuevo brote vírico o si por el contrario se trata de un caso excepcional.

El resto de los datos de cada tuit no son necesarios para la creación de los indicadores que permitan detectar más precozmente la aparición de un brote vírico. Estos datos descartados podrían ser tenidos en cuenta para futuros análisis, como por ejemplo encontrar a aquellas personas que más hablan sobre la gripe y que tienen una mayor repercusión en la transmisión de mensajes relacionados. Detectar a las personas más influyentes nos puede servir para cumplir mejor algunos objetivos estratégicos como puede ser ofrecer una mejor información de carácter preventivo a los ciudadanos a través de personas relevantes que hablen a menudo sobre la gripe en Twitter.

3.3.2. Fichero de hospitalizaciones

El fichero de hospitalizaciones contiene información sobre los casos de hospitalizaciones que tuvieron lugar en el ICS entre enero de 2012 y marzo de 2013.

Cada entrada del conjunto de datos contiene los siguientes atributos:

Tabla 4. Atributos del conjunto de los datos sobre hospitalizaciones

Nombre del campo	Descripción
Any Alta	Año de alta
Clase Admisió	Tipo admisión. 1: urgente; 2: programada
Clase Admisio Hospitalitzacio Posterior	Igual que anterior, si se ha producido
Clase Alta	Tipo alta: 1: a domicilio; 2: agudos/psiquiátrico; 3: Instituto Catalán de Oncología; 4: socio sanitario; 4: residencia social; 5: voluntaria; 6: exitus; 7: huida del paciente; 8: H. domiciliaria; N/D: N/D
Data Ingrés	Fecha de ingreso
Data Alta	Fecha de alta
Data Episodi Urg Seg	Si ha habido urgencia posterior
Data Hospitalitzacio Anterior	Si es reingreso
Data Hospitalitzacio Posterior	Número episodio asistencia: identifica unívocamente
Es Acumulat Any Actual Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Acumulat Any Anterior Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Any Anterior Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Actual Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Actual Any Anterior Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Corrent Any Actual Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Corrent Any Anterior Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Setmana Actual Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Setmana Any Anterior Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Setmana En Curs Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Grd Codigo	Código GRD
GRD Descripcion	Descripción código GRD
Nhc	Numero historia clínica pacientes (están modificados)
Periode Alta	Periodo alta (mes/año). Dato auxiliar
Servei Hospitalitzacio	Código Servicio hospitalización al alta
Servei_descriptiu	Descripción código
Sexe	0: hombre; 1: mujer

Nombre del campo	Descripción
Tipus Episodi	H: hospitalización: CM: cirugía mayor ambulatoria; DO: domiciliaria
Dies Hospitalitzacio	Días totales de hospitalización
Estada en Rea_PQ (h)	Estancia en servicios especiales (en milisegundos)
Estada en UCI	Estancia en servicios especiales (en milisegundos)
Estada en UCI CCA	Estancia en servicios especiales (en milisegundos)
Estada en Unitat Coronaria	Estancia en servicios especiales (en milisegundos)
Numero unitats especials ha passat	Servicios especiales por los que ha pasado
Peso GRD	Complejidad clínica, peso del GRD para valorar la
Temps Hospitalitzacio (h)	Tiempo hospitalización (en horas)
Diagnostic P (codi)	Diagnóstico al alta
Diagnostic P (desc)	Descripción código alta

De la misma manera que sucedía con los datos sobre los tuits, de entre todos los datos disponibles en el fichero sobre los casos de hospitalización, solo son útiles una parte de ellos para la elaboración de indicadores para la detección precoz de brotes gripales. Entre todos los datos se utilizarán la fecha de admisión, la fecha de alta, la descripción de la patología (GRD descripción), el servicio descriptivo (o unidad hospitalaria) donde fue tratado el paciente y el sexo del paciente.

La descripción de la patología (GRD descripción) es muy importante porque es en este campo donde tenemos la información sobre si la hospitalización está relacionada con la gripe. La fecha de hospitalización es necesaria para cruzar los datos de las hospitalizaciones por gripe con los indicadores propuestos a partir del número de mensajes escritos por semana en Twitter. Al cruzar la información de ambas fuentes de datos podremos validar si los indicadores propuestos tienen realmente capacidad predictiva y son útiles para la detección precoz de brotes víricos de gripe.

La fecha de alta, la unidad hospitalaria donde fue tratado el paciente y su sexo servirán para cumplir con el objetivo estratégico de realizar una gestión más eficiente de los recursos sanitarios disponibles, tanto de personal como de material médico.

En cambio, el resto de los datos disponibles en el fichero de datos de hospitalizaciones, si bien pueden ser muy útiles en otros ámbitos de control, no son de importancia para la detección precoz de brotes víricos de gripe.

3.3.3. Fichero de casos de urgencias

El fichero de casos de urgencias contiene información sobre los casos de urgencias que tuvieron lugar en el ICS entre los meses de enero a marzo de los años 2012 y 2013.

Cada entrada del conjunto de los datos contiene los siguientes atributos:

Tabla 5. Atributos del conjunto de datos sobre los casos de urgencias

Nombre del campo	Descripción
ABS	Área de referencia del paciente, para identificar pacientes fuera de zona
Abs Desc Sap	Descripción anterior
Abs Hospital Referencia	Tipo admisión si hay hospitalización previa: 1: urgente; 2: programada
Clase Admisio Episodi Hospitalitzacio Anterior	Tipo admisión si hay hospitalización previa: 1: urgente; 2: programada
Clase Admisio Episodi Hospitalitzacio Posterior	Tipo admisión si hay hospitalización posterior: 1: urgente; 2: programada
Clase Alta	Tipo alta: 1: a domicilio; 2: agudos/psiquiátrico; 3: Instituto Catalán de Oncología; 4: Socio sanitario; 4: residencia social; 5: voluntaria; 6: exitus; 7: huida del paciente; 8: H. domiciliaria; N/D: N/D
Data Alta Complet	Fecha alta: valores auxiliares
Data Alta Dia Setmana Literal	Fecha alta: valores auxiliares
Data Alta Periode	Fecha alta: valores auxiliares
Data Entrada	Fecha alta: valores auxiliares
Data Episodi Hospitalitzacio Anterior	Fecha episodio hospitalización anterior, si procede
Data Episodi Hospitalitzacio Posterior	Fecha episodio hospitalización posterior, si procede
Data Episodi Urgencies Anterior	Fecha episodio urgencias anterior, si procede (a efecto de identificar readmisiones)
Data Episodi Urgencies Posterior	Fecha episodio urgencias posterior, si procede (a efecto de identificar readmisiones)
Data Hora Assistencia	Hora inicio asistencia médica
Data Hora Entrada	Hora inicio entrada administrativa
Data Hora Episodi Anterior	Hora entrada episodio anterior, para calcular readmisiones
Data Hora Sortida	Hora salida
Data Neixement	Fecha nacimiento, a efectos de calcular edad pacientes
Data Sortida	Fecha salida
Episodi	Código identificación episodio actual
Episodi Hospitalitzacio Anterior	Código identificación hospitalización anterior

Nombre del campo	Descripción
Episodi Hospitalitzacio Posterior	Código identificación hospitalización posterior
Episodi Urgencies Anterior	Código episodio urgencias anterior
Episodi Urgencies Posterior	Código identificación episodio actual
Es Acumulat Any Actual Data Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Acumulat Any Anterior Data Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Any Anterior Data Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Actual Any Actual Data Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Es Mes Actual Any anterior Data Alta	Valores auxiliares para identificar periodo análisis: para realizar comparativas anuales, mensuales, semanales, etc.
Hora entrada	Hora de entrada: Texto, formato HHMMSS
Hora Sortida	Hora salida: Texto, formato HHMMSS
NHC	Identificador del paciente
Pais Residencia	País de residencia, para identificar no residentes
Servei Alta	Servicio de alta
Servei alta Descriptiu	Descripción anterior
Servei Alta Tipus Unitat Organitzativa	Tipo servicio: generales, médicos, quirúrgicos, pediatría
Servei Entrada	Servicio asignado al ingreso
Sexe	0: hombre; 1: mujer
Triatge Darrer	Nivel de triaje último: 000, 100, 200, 300, 400, 500 de mayor criticidad a menor. No hay triaje a todas horas
Triatge Primer	Nivel de urgencias: 000, 100, 200, 300, 400, 500 de mayor criticidad a menor. No hay triaje a todas horas
Minuts a Urgencies	Tiempo en urgencias

En el fichero de datos de urgencias no hay datos disponibles sobre la gripe, por tanto se ha decidido no utilizar los datos de este fichero.

3.3.4. Fuentes de datos externas

Además de los tuits sobre la gripe recuperados a partir de la API de Twitter, también es posible la utilización de fuentes de datos externas para la elaboración de indicadores.

El mejor ejemplo de utilización de fuentes de datos externa es Google Flu Trends, un proyecto de Google que muestra el recuento de búsquedas sobre la gripe que realizan los usuarios de Google en la mayoría de los países de Europa,

América y Oceanía. Una de las ventajas de Google Flu Trends es que pone a nuestra disposición el histórico de búsquedas sobre la gripe realizadas por país en Google desde el año 2004. En el ámbito del territorio español, Google Flu Trends ofrece las búsquedas realizadas en cada comunidad autónoma.

Google Flu Trends

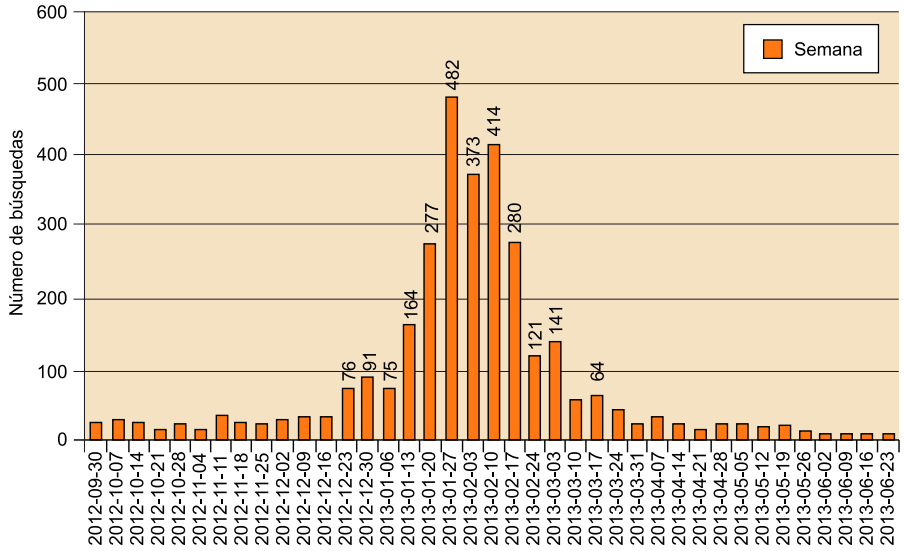
En este enlace está disponible el histórico del número de búsquedas por semana sobre la gripe en un gran número de países. Y en este otro enlace está disponible el histórico del número de consultas realizadas sobre la gripe en cada comunidad autónoma de España.

Además Google Flu Trends también permite visualizar en un mapa del mundo el histórico de búsquedas por fechas para cada país y región, indicando también la probabilidad de que cada país o región esté en un periodo de brote de gripe.

Por el contrario, el proyecto no indica las palabras clave utilizadas para hacer el recuento de las búsquedas sobre la gripe. Además, la única información disponible es el número de búsquedas realizadas y, si bien es muy útil para la creación de indicadores que permitan predecir la aparición de un brote de gripe, no aporta ninguna información extra que pueda ser utilizada para cumplir otros objetivos estratégicos del cuadro de mando.

En cualquier caso, dado que tanto Twitter como Google reflejan la interacción de los usuarios con Internet, las gráficas que muestran uno y otro deberían tener una forma similar. El siguiente gráfico muestra el número de consultas sobre la gripe realizadas en Google en el área de influencia del ICS. Podemos observar que el gráfico tiene una forma similar al gráfico sobre los tuits escritos sobre la gripe (ver figura 5) y las fechas de incremento y decremento del número de mensajes coinciden.

Figura 4. Búsquedas en Google sobre la gripe por semana realizadas en Cataluña



Para elaborar el gráfico hemos recuperado el número de búsquedas sobre la gripe realizadas por semana en Cataluña y hemos creado una fuente de datos que contendrá los campos de la fecha de inicio de la semana y el número de consultas sobre la gripe realizadas en esa semana. Estos datos, aunque vienen de una fuente externa no prevista inicialmente, aportarán mayor robustez a los indicadores creados a partir de los mensajes escritos en Twitter y permitirán evaluar la correlación entre mensajes escritos en Twitter y búsquedas realizadas en Google.

La fuente de datos contiene el número de búsquedas realizadas sobre la gripe durante el año 2013 en cada una de las 52 semanas del año.

Tabla 6

Nombre del campo	Descripción
Semana	Este campo contiene la fecha del primer día de la semana para la que se contabilizan las búsquedas
Número de búsquedas	Número de búsquedas sobre la gripe realizadas en Google en la región de Cataluña en la semana en cuestión

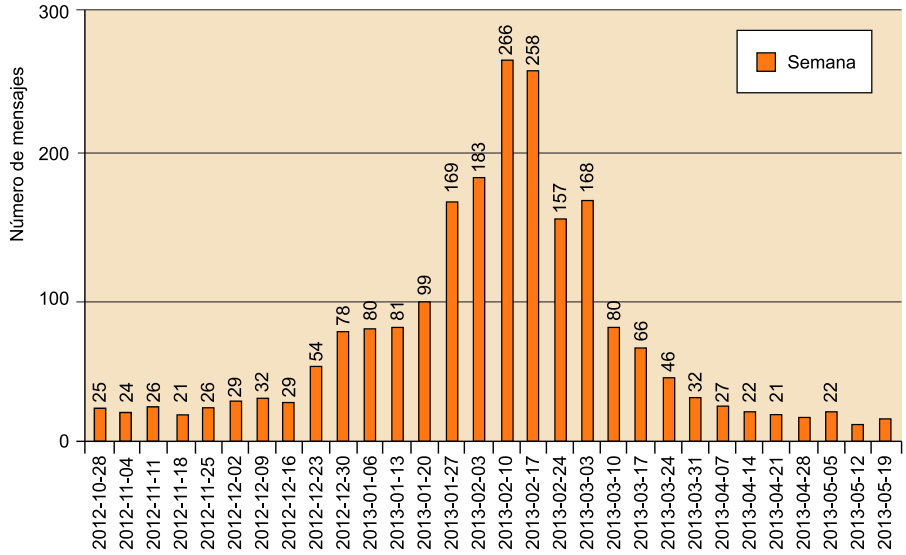
3.4. Indicadores para la predicción de brotes de gripe

Antes de elaborar el cuadro de mando debemos analizar las fuentes de información disponibles y los indicadores que contienen. Muchos de estos indicadores, como pueden ser el número de altas totales o el número de defunciones, aportan información directa que puede ser integrada fácilmente en un cuadro de mando. En cambio la información que obtenemos de Twitter puede no ser tan fácil de interpretar. En este caso, antes de definir un indicador se deberá cruzar la información obtenida de Twitter con los datos hospitalarios disponibles para validar que los KPI propuestos son válidos para conseguir los objetivos planteados.

El primer paso consistirá en obtener el número de mensajes escritos por los usuarios cada semana durante el periodo de estudio a partir del fichero de tuits.

En la figura 1 podemos ver la evolución semanal del número de mensajes escritos por los usuarios durante un periodo de 6 meses entre los meses de noviembre de 2012 y junio de 2013.

Figura 5. Número de tuits por semana sobre la gripe escritos por los usuarios de Twitter



A partir de la gráfica podríamos definir algunos indicadores que podrían ser útiles para predecir un nuevo brote vírico de gripe:

1) Que el número de mensajes escritos por los usuarios sea superior a *x*.

Un indicador que utilice un umbral numérico fijo no es una buena solución porque el número de mensajes escritos por los usuarios puede variar sustancialmente de un año a otro, ya sea por un incremento de los usuarios de Twitter, por una mayor actividad de los usuarios, por un decremento de los usuarios o por la cantidad de mensajes escritos por ellos.

2) Variación porcentual semanal en el número de mensajes escritos

Un indicador de este tipo podría ser un buen indicador ya que permite incrementos significativos en el número de mensajes escritos sobre la gripe. En el gráfico podemos observar que entre la semana del 16 al 23 de diciembre de 2012 se produce un incremento porcentual del 86,2%[3], entre la semana del 23 al 30 de diciembre se produce un incremento del 44,4%[4] y entre la semana del 20 al 27 de enero de 2013 se produce un incremento del 70,7%[5].

[3] ((54/29) * 100) = 86,2%.

[4] ((78/54) * 100) = 44,4%.

[5] ((169/99) * 100) = 70,7%.

Estos cambios porcentuales podrían ser un buen indicador para predecir la llegada o el fin de un nuevo brote vírico. No obstante, también deberemos tener en cuenta que puede haber cambios significativos en el número de mensajes escritos que no sean indicadores de la llegada de un nuevo brote vírico. Un incremento significativo en el número de mensajes puede darse debido a la aparición de una noticia en un medio de difusión masivo, como por ejemplo que un futbolista popular se pierda un partido importante debido a una gripe, que se produzca un avance importante en una vacuna para la gripe, que

se den casos similares a la gripe aviar hace unos años o que salga un nuevo sistema similar al que estamos definiendo en este caso práctico para detectar con mayor antelación la llegada de un brote vírico.

Además, también pueden producirse cambios porcentuales significativos en épocas que no preceden a un virus en las que se escriben muchos menos mensajes sobre la gripe. Debido al reducido número de mensajes escritos sobre la gripe, puede haber incrementos superiores al 50% en el número de mensajes escritos sobre la gripe entre dos semanas sin que un nuevo brote de gripe se produzca en el corto plazo.

3) Variación porcentual en dos semanas en el número de mensajes escritos

La variación porcentual en el número de mensajes escritos en un periodo de 15 días podría ser un buen indicador ya que tendría las mismas ventajas que el indicador que evalúa la variación porcentual semanal en el número de mensajes escritos que acabamos describir, pero minimizaría el impacto que podría tener una noticia de relevancia.

4) Combinación de variaciones porcentuales y del número de mensajes escritos

Para minimizar las carencias de los indicadores que se acaban de analizar, se podría utilizar la combinación de un umbral mínimo de mensajes por semana y el incremento porcentual en el número de mensajes escritos.

Antes de decidirnos por uno o varios KPI, es aconsejable cruzar el número de mensajes escritos por semana en Twitter con los datos disponibles en los ficheros de urgencias y hospitalizaciones. El segundo paso consistirá en validar si los posibles indicadores que hemos propuesto se corresponden a periodos de brotes de gripe.

Analizando el fichero de urgencias, podemos observar que disponemos de información sobre el departamento del hospital en el que fue atendido el paciente, pero no tenemos información detallada para saber si el usuario llegó a urgencias por un caso de gripe.

En cambio, en el fichero de hospitalizaciones disponemos de la descripción de la patología sufrida por el paciente y encontramos un total de 14 hospitalizaciones por casos de gripe que resumimos en la siguiente tabla:

Tabla 7. Resumen de hospitalizaciones por casos de gripe

Fecha de admisión	Fecha de alta	GRD descripción	Servicio descriptivo	Sexo
21/01/12 17:34	23/01/12 13:53	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H

Fecha de admisión	Fecha de alta	GRD descripción	Servicio descriptivo	Sexo
25/01/12 11:16	26/01/12 13:52	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
02/02/12 14:25	23/02/12 18:33	Gripe causada por virus identificado de la gripe aviar con neumonía	Neumología	H
05/02/12 01:00	05/02/12 18:55	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
11/02/12 02:28	13/02/12 14:51	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
23/02/12 16:47	27/02/12 14:14	Gripe con neumonía	Pediatría	H
27/02/12 07:54	28/02/12 14:59	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
22/01/12 10:45	22/02/12 07:53	Gripe causada por virus identificado de la gripe aviar con neumonía	Medicina intensiva	H
24/01/13 23:15	27/01/13 14:58	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
27/01/13 20:27	29/01/13 11:36	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
28/01/13 18:49	30/01/13 12:15	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H
28/01/13 20:33	03/02/13 20:10	Gripe con neumonía	Neumología	H
21/02/2013 13:00:38	27/02/13 14:23	Gripe con neumonía	Neumología	H
22/02/2013 12:47:54	27/02/13 11:08	Gripe con otras manifestaciones respiratorias. Gripe NOS, gripal: laringitis, faringitis, infección respiratoria	Pediatría	H

De las 14 hospitalizaciones, 10 casos se trataron en la unidad de pediatría, 1 caso en la unidad de medicina intensiva y 3 casos en la unidad de neumología.

Las hospitalizaciones por gripe tienen lugar tanto en el año 2012 como en el 2013 y en ambos casos las primeras hospitalizaciones se producen a finales de enero y las últimas a finales de febrero. Debido a que no disponemos del histórico de tuits de inicios de 2012, deberemos centrar el análisis en los casos de hospitalización del año 2013.

Si comparamos el periodo en el que se produjeron las hospitalizaciones por casos de gripe con el número de mensajes escritos en Twitter por semana, vemos que las hospitalizaciones tuvieron lugar en el periodo en que se escribieron más mensajes sobre la gripe en la red social.

Sin embargo, el KPI que escojamos debe ayudarnos a predecir con antelación la aparición o el fin de un brote de gripe. Por tanto, nuestro KPI deberá ser capaz de indicarnos la llegada o finalización de un nuevo brote vírico con 2 o 3 semanas de antelación. Esto permitirá prevenir al ciudadano con mayor antelación y gestionar mejor los recursos disponibles para tratar los casos de hospitalizaciones y urgencias por gripe.

A partir del **cruce de datos** de hospitalizaciones y el número de tuits escritos sobre la gripe, como **KPI para la detección de un brote vírico de gripe** escogeremos una combinación de un **umbral mínimo de 75 mensajes escritos en una semana** y un **incremento porcentual superior al 200% en un plazo de dos semanas**.

Con esta combinación intentamos eliminar el riesgo de la toma de decisiones si el número de mensajes escritos no es significativo y el riesgo de una variación porcentual de mensajes escritos debido a la aparición de una noticia sobre la gripe que no esté relacionada con la aparición de un nuevo brote vírico.

Para finalizar, debemos tener en cuenta que estos KPI han sido escogidos a partir de la información disponible de solo un año, así que en un futuro deberán validarse o actualizarse a partir de nuevos datos disponibles.

3.5. Contenido gráfico del cuadro de mando

Una vez definidos los indicadores que vamos a utilizar, se definirá un esbozo del cuadro de mando de forma gráfica para que los gestores del ICS puedan detectar de forma más precoz la llegada de un brote de gripe.

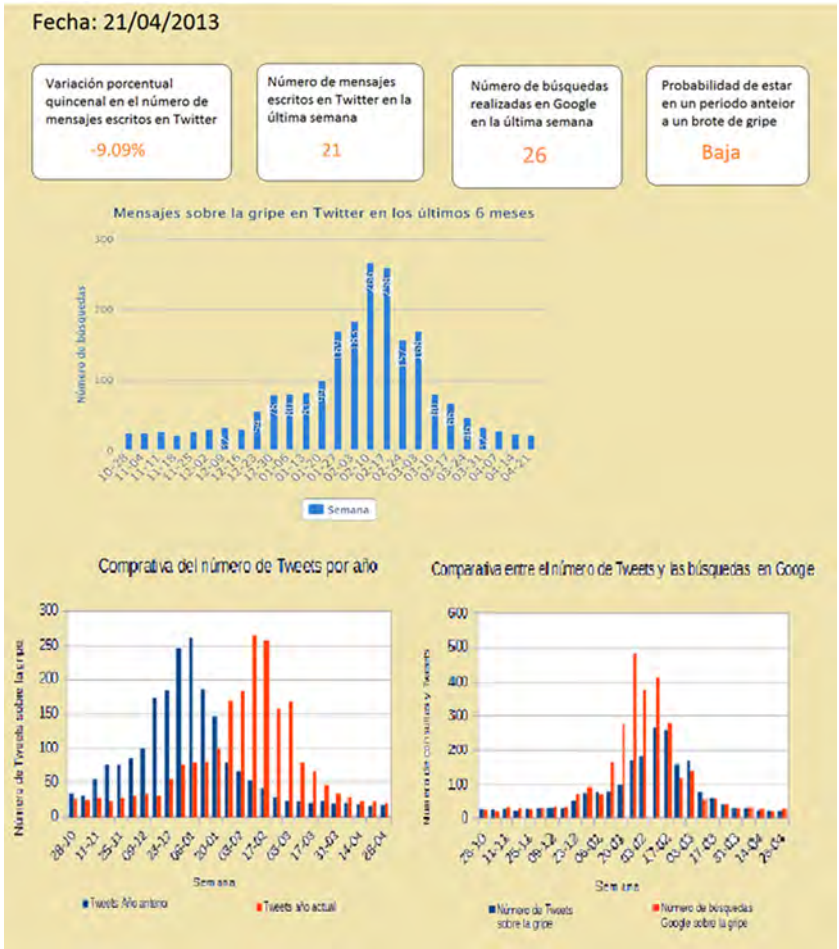
El cuadro de mandos contará con los indicadores propuestos sobre la **variación quincenal del número de tuits sobre la gripe** y el **número de tuits escritos durante la semana**, así como el **número de búsquedas sobre la gripe realizadas en Google** y una serie de gráficos que permiten **comparar la evolución del número de mensajes escritos el año anterior con el número de mensajes escritos este año**, el **número de mensajes sobre la gripe escritos en**

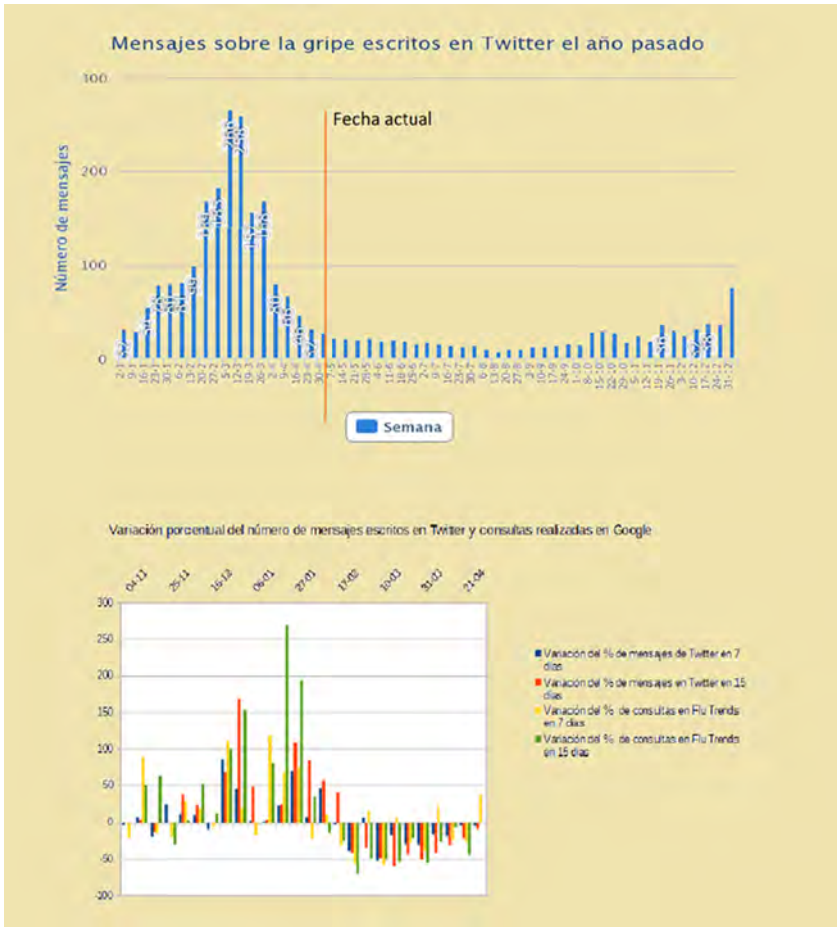
Twitter con el número de búsquedas realizadas en Google y un último gráfico para ver las **variaciones porcentuales en el número de mensajes escritos en Twitter** y **consultas realizadas en Google en intervalos de 7 y 15 días**.

Con estas gráficas se permitirá evaluar la evolución del número de mensajes escritos en Twitter y sus incrementos porcentuales respecto a los últimos 7 y 15 días, comparar los datos del año actual con los del año anterior y comprar la información de Twitter con la información ofrecida por fuentes externas como la del proyecto Google Flu Trends.

Para facilitar la interpretación de la situación actual, el cuadro de mando contendrá un indicador que mostrará si la situación actual es una época sin riesgo de llegada de un brote vírico o si estamos en un posible periodo anterior a un brote gripal. Este indicador estimará si estamos en un periodo con alta probabilidad de preceder a un brote de gripe. Para hacerlo se comprueba si el número de mensajes de Twitter en la presente semana es superior a 75 y el incremento porcentual del número de mensajes escritos en Twitter en las últimas dos semanas es superior al 200%. Cuando esto pasa entenderemos que estamos en un periodo de probabilidad alta. En caso contrario consideraremos que estamos en un periodo de baja probabilidad de preceder a un periodo gripal.

Figura 6. Cuadro de mandos





Se debe indicar que el cuadro de mando únicamente contiene información para la detección precoz de nuevos brotes de gripe. En un entorno real probablemente el cuadro de mando no contendría solamente información para la detección de nuevos brotes gripales, sino que podría contener otros indicadores relativos a otros ámbitos de la gestión hospitalaria. Muchos de estos indicadores podrán agruparse juntos en una misma interfaz gráfica (número de ingresos, de altas, de defunciones, etc.), mientras que otros más específicos, como es la detección de brotes gripales, tendrían su propia sección específica dentro del cuadro de mando.

4. Diseño e implementación del almacén de datos

En este capítulo se diseñará e implementará el almacén de datos que dé soporte al cuadro de mando que hay que implementar. Primero se analizarán las necesidades de almacenamiento y se escogerá que tipo de sistema gestor de base de datos se utilizará. Posteriormente se realizará un diseño conceptual del almacén de datos para al final implementarlo.

4.1. Necesidades de almacenamiento

Para la predicción de brotes de gripe a partir de los mensajes escritos por los usuarios de Twitter se utilizan datos de los tuits sobre la gripe y el número de búsquedas sobre la gripe realizadas por los usuarios en Google. También se han utilizado datos sobre hospitalizaciones para validar que el incremento en el número de mensajes escritos en Twitter y en las búsquedas realizadas en Google están correlacionados con el aumento de hospitalizaciones por gripe. Aunque los datos de hospitalizaciones no se utilizan para calcular los indicadores diseñados, hemos decidido añadirlos en el almacén de datos por completitud y para permitir futuras evoluciones.

El primer paso para definir el sistema de almacenamiento consistirá en realizar una estimación del tamaño máximo que podrá alcanzar nuestro sistema de almacenamiento. El resultado de dicho estudio condicionará el tipo de sistema de bases de datos que se elegirá.

4.1.1. Estimación del tamaño máximo de almacenamiento

1) Tuits escritos relacionados con la gripe

Para hacer esta estimación utilizaremos el conjunto de datos de Twitter que contiene los mensajes escritos sobre la gripe entre noviembre de 2012 y abril del 2013.

El número de campos que se almacenará para cada tuit en el almacén de datos será reducido: el nombre del usuario, el enlace al perfil del autor del tuit y la fecha en la que fue escrito. Por tanto el tamaño requerido por cada tuit será muy bajo.

En los 6 meses que incluyen el periodo de máxima actividad de la gripe se han generado un total de 2.102 registros. Teniendo en cuenta que en las semanas de baja actividad de gripe se escriben en Twitter unos 22 mensajes de media sobre la gripe y que el fichero de datos contiene los 6 meses en los que tiene lugar el brote de gripe, podemos estimar que la cantidad aproximada de tuits adicionales que se almacenarán será de 528 tuits, que corresponderá a los 22

tuits por semana de cada una de las 24 semanas restantes por almacenar. Por tanto, se estima que se almacenarán alrededor de 2.600 tuits anuales sobre la gripe para el área de actuación del ICS, un número tampoco demasiado elevado en el contexto de los almacenes de datos.

2) Hospitalizaciones

El fichero de datos sobre las hospitalizaciones contiene los datos sobre las altas hospitalarias en el periodo de enero a marzo de 2012 y de 2013. En la siguiente tabla se analiza el número de altas por mes en el periodo del conjunto de datos:

Tabla 8. Número de altas por mes en el periodo del conjunto de datos

Mes	Número de altas	Número de días	Media de altas por día
Enero 2012	2.287	31	73,61
Febrero 2012	2.557	29	88,17
Marzo 2012	2.598	31	83,81
Primer trimestre 2012	7.442	91	81,78
Enero 2013	2.327	31	75,06
Febrero 2013	2.282	28	81,5
Marzo 2013	2.269	31	73,19
Primer trimestre 2013	6.878	90	76,42

Como podemos observar en la tabla anterior, los datos de altas hospitalarias alcanzan un máximo de 2.600 altas por mes y la media de altas por día alcanza un valor máximo de 89 altas de media por día. Si cogemos el valor de altas hospitalarias más elevado y le aplicamos un margen de seguridad, podemos prever el número máximo de registros que necesitará nuestro sistema de almacenamiento para guardar toda la información de las altas hospitalarias de un año. Tomando un valor máximo de 95 altas hospitalarias por día obtenemos un valor máximo de 34.675 registros anuales.

No obstante, no todos los 34.675 registros deben almacenarse en el almacén de datos. Solo deberán almacenarse aquellos que traten sobre la gripe. Tal y como hemos comentado, para las fechas analizadas (los dos meses de más actividad en 2012 y 2013) hemos encontrado solamente 14 hospitalizaciones relacionadas con la gripe. A la vista de esa información, podríamos estimar que no se esperan más de 50 hospitalizaciones anuales por casos de gripe. Por tanto, podríamos concluir que las necesidades de almacenamiento en este caso serían mínimas.

3) Búsquedas en Google sobre la gripe

Solo disponemos del número de consultas sobre gripe realizadas por semana. Por tanto, el número de registros anuales que se almacenarán serán 52, uno por semana. La información que se almacenará será mínima: la fecha y el número de consultas realizadas.

Sumando las necesidades de almacenamiento de cada uno de los conjuntos de datos, tenemos unas necesidades de almacenamiento anuales de 2.752 registros, donde cada uno de ellos requiere de muy pocos datos. Por tanto, los requisitos de almacenamiento son mínimos y no deberán condicionar el modelo de datos escogido para implementar el almacén de datos.

Tabla 9. Parámetros para realizar la consulta

Fuentes de datos	Máximo
Twitter	2.600
Hospitalizaciones	100
Google	52
Total	**2.752**

4.1.2. Elección del modelo de datos

Tal y como hemos justificado en el punto anterior, las necesidades de almacenamiento son mínimas, por tanto no será necesario un sistema de almacenamiento que permita escalabilidad horizontal. Por otro lado, el sistema que se creará no será un sistema en tiempo real, sino que los datos del almacén de datos se cargarán de forma periódica (por ejemplo, los tuits podrían cargarse diariamente y la información de búsquedas en Google semanalmente).

Con estas características tenemos que decidir si el sistema de almacenamiento que mejor se adapta a nuestras necesidades responde a un modelo de datos NoSQL o relacional. En particular, los modelos que se barajarán serán los modelos de grafo, clave-valor, documental, agregado en columnas y relacional.

Hemos considerado el modelo de grafo inadecuado para el problema que se trata. El motivo es que los datos que se almacenarán no contienen gran cantidad de relaciones entre ellos, ni la dificultad en los cálculos responde al análisis de complejas relaciones entre los datos.

Respecto a los modelos agregados (clave-valor, documental y columnas), hay que decir que el primero y el tercero no son especialmente adecuados para este caso.

Consideramos el modelo clave-valor inadecuado por diversos motivos. Su modelo no permite almacenar los datos de forma estructurada y eso puede llevar a complicaciones a la hora de relacionar los datos y a calcular los datos en distintos niveles de abstracción (tuits por días, por semanas y por meses, por

ejemplo). Las claves que se utilizan (el número de hospitalizaciones o el identificador de un tuit) no son claves fácilmente reconocibles ya que no aportan ningún tipo de información semántica para reconocer un caso en concreto. Por otro lado, al no ser necesaria una alta disponibilidad ni la posibilidad de escalar horizontalmente, las ventajas que ofrecen estos modelos de datos no son aprovechadas.

El modelo de datos de agregación en columnas ha sido descartado porque el número de registros que se almacena y el tamaño son muy pequeños, y estos sistemas están optimizados para tratar gran cantidad de datos y de gran tamaño.

Las bases de datos NoSQL documentales podrían ser una buena opción de almacenamiento en este caso. Es cierto que, como se ha dicho antes, algunas claves que se utilizan no son muy intuitivas y pueden llevar a problemas o confusiones, pero los datos requeridos para calcular los indicadores (número de tuits por semana, número de búsquedas por semana, incremento de tuits por semana, etcétera) son pocos y reutilizables entre indicadores. Eso permitiría precalcular estos datos en distintos agregados durante los procesos ETL y guardarlos como agregados, lo que permite calcular los indicadores de forma rápida y sencilla. Esto sería viable debido a la baja cantidad de registros y al poco tamaño de estos. No obstante, este tipo de base de datos dificultaría la creación de nuevos indicadores, consultas y/o filtros de datos.

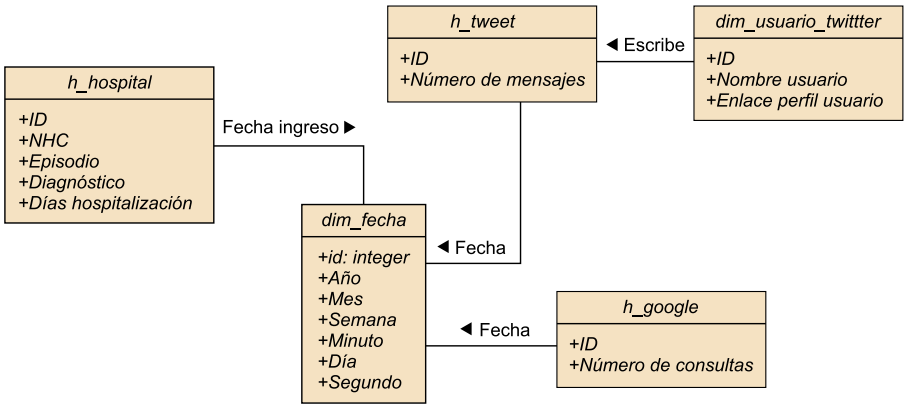
Otra opción viable sería el uso de un modelo relacional clásico. Este modelo sería aplicable debido a su gran personalización, a que permite realizar consultas complejas para recuperar información histórica sobre la base de distintos parámetros y permite realizar consultas ad hoc que no se hayan especificado a priori, que podrían ser útiles para validar si los indicadores son lo suficientemente predictivos o es mejor utilizar unos nuevos.

En conclusión, para el caso de estudio tanto una base de datos documental como una base de datos relacional serían adecuadas. En este caso los autores se han decantado por utilizar una base de datos relacional debido a su madurez y a la alta integración que tienen con las suites de inteligencia de negocio.

4.2. Diseño del almacén de datos

A partir del análisis realizado hasta el momento se han identificado tres tablas de hechos: una para los tuits, otra para las búsquedas de Google y otra para los datos de hospitalizaciones. Estas tres tablas de hechos estarán relacionadas en el almacén de datos mediante la fecha, que es el nexo en común de las tres. El siguiente diagrama muestra el diseño conceptual del almacén de datos en UML:

Figura 7. Diseño conceptual del almacén de datos en UML



A continuación presentamos cada una de estas tablas de hechos junto con sus tablas de dimensiones. Finalmente presentaremos el modelo lógico creado para adaptar el diseño del almacén de datos al modelo relacional.

4.2.1. Datos de Twitter

Para el conjunto de datos de Twitter se ha identificado una tabla de hecho y dos dimensiones (usuario y fecha).

Los campos "palabra clave" e "idioma" disponibles en el conjunto de datos de Twitter podrían haberse considerado, pero se han descartado debido a que no se han considerado relevantes para el problema que se trata. El contenido del tuit, al no aportar información relevante para la generación de los indicadores, tampoco se almacenará en la tabla de hechos.

La tabla de hecho correspondiente a los datos recuperados de Twitter es:

Tabla 10. Tabla de hecho de los datos recuperados de Twitter

Tabla de hecho	Descripción
h_tweet	Recoge los mensajes de Twitter que tratan sobre la gripe

La tabla de hechos contendrá el número de tuits semanales que tratan sobre la gripe por usuario.

Las dimensiones corresponden a las perspectivas de negocio sobre las que queremos analizar el conjunto de mensajes de tuits:

Tabla 11. Tabla de dimensiones

Dimensiones	Descripción
d_fecha	Fecha en que se ha escrito el tuit
d_usuarioTwitter	Usuario que ha escrito el tuit

4.2.2. Datos de búsquedas en Google

Respecto al conjunto de datos de Google se han identificado una tabla de hechos y una tabla de dimensión.

Tabla 12. Tabla de hecho de los datos recuperados de Google

Tabla de hecho	Descripción
h_google	Recoge las consultas sobre la gripe realizadas en Google

La tabla de hecho contendrá el número de consultas semanales sobre la gripe realizadas en Google.

La dimensión propuesta es la correspondiente a la semana en la que se han realizado las búsquedas.

Tabla 13. Tabla de dimensiones

Dimensiones	Descripción
d_fecha	Semana en la que se han realizado las búsquedas

4.2.3. Datos de hospitalizaciones

Respecto a los datos de hospitalizaciones se ha definido una tabla de hechos y una dimensión.

Tabla 14. Tabla de hecho de los datos de hospitalizaciones

Tabla de hecho	Descripción
h_hosp	Recoge el proceso de hospitalización de un paciente

El conjunto de datos del fichero de hospitalizaciones nos interesa para validar que los indicadores propuestos para la detección precoz de nuevos brotes de gripe son realmente eficaces, y para ello cruzaremos la información obtenida de los indicadores con los casos de ingresos hospitalarios causados por la gripe.

De todo el conjunto de datos disponibles en el fichero de hospitalizaciones, solo almacenaremos aquellos datos que nos permitan identificar el hecho, conocer que la hospitalización fue causada por la gripe y el tiempo que duró la hospitalización. Los campos seleccionados del fichero de hospitalizaciones son:

- *Episodio*: Número de episodio de asistencia: identifica unívocamente.
- *NHC*: El número de historia clínica del paciente.
- *Diagnóstico P* (descripción): descripción de la patología del enfermo.
- *Fecha de ingreso*.

- *Días de hospitalización.*

La dimensión *fecha* nos permitirá indicar la fecha de ingreso del paciente.

Tabla 15. Tabla de dimensiones

Dimensiones	Descripción
d_f_ingreso	Fecha de ingreso

4.2.4. Diseño lógico

Una vez determinados qué hechos, dimensiones, métricas y atributos existen podemos determinar el diseño lógico de la base de datos que se creará. Al utilizar un modelo relacional se deberán identificar las distintas tablas, sus atributos y las claves foráneas que representaran las relaciones entre los datos.

La tabla de hechos de tuits se llamará *H_TWEET* y estará formada por los siguientes elementos:

Tabla 16. Tabla *H_TWEET*

	h_tweet	**Tipo**
(PK)	*id_tweet*	Clave primaria
(FK)	*id_fecha*, *id_usuarioTwitter*	Claves foráneas
	NumeroMensajes	Métrica

La tabla de hechos de búsquedas en Google se llamará *H_GOOGLE* y estará formada por los siguientes elementos:

Tabla 17. Tabla *H_GOOGLE*

	h_google	**Tipo**
(PK)	*id_google*	Clave primaria
(FK)	*id_semana*	Clave foránea
	NumeroBusquedas	Métrica

La tabla de hechos de hospitalizaciones se llamará *H_HOSP* y estará formada por los siguientes elementos:

Tabla 18. Tabla *H_HOSP*

	H_hosp	**Tipo**
(PK)	*id_hosp*	Clave primaria
(FK)	*id_d_f_ingreso*	Clave foránea
	d_hosp	Métrica

	H_hosp	Tipo
	Episodio *NHC*	Atributos

Para representar la dimensión *fecha* de cada una de las tablas de hechos podríamos crear tablas distintas (fecha de tuit, fecha de Google y fecha de hospitalización) o utilizar una sola tabla para representar todas las fechas. No tiene sentido crear diferentes tablas para representar los mismos datos. Por tanto se ha decidido almacenar las fechas en una única tabla y, por tanto, todas las claves foráneas que relacionen fechas apuntarán a la misma tabla. La dimensión *fecha* contendrá algunos atributos con valores nulos para algunas fechas, ya que las fechas de semana, no contendrán la hora, el minuto y el segundo, pero todas ellas contendrán el año, el mes y el día. De esta forma simplificaremos el modelo final con el que trabajamos y el proceso de carga de dimensiones. Las dimensiones resultantes serán:

Tabla 19. Tabla de dimensiones

Dimensiones	Atributos
d_fecha	*Año, mes, semana, día, hora, minuto, segundo*
d_usuario_twitter	*NombreUsuario, EnlacePerfilUsuario*

4.3. Implementación del almacén de datos

Para implementar el almacén de datos se ha decidido utilizar Oracle Express Edition 11g.

Oracle Express Edition 11g

Podéis descargar el instalador de Oracle Express Edition en este enlace.

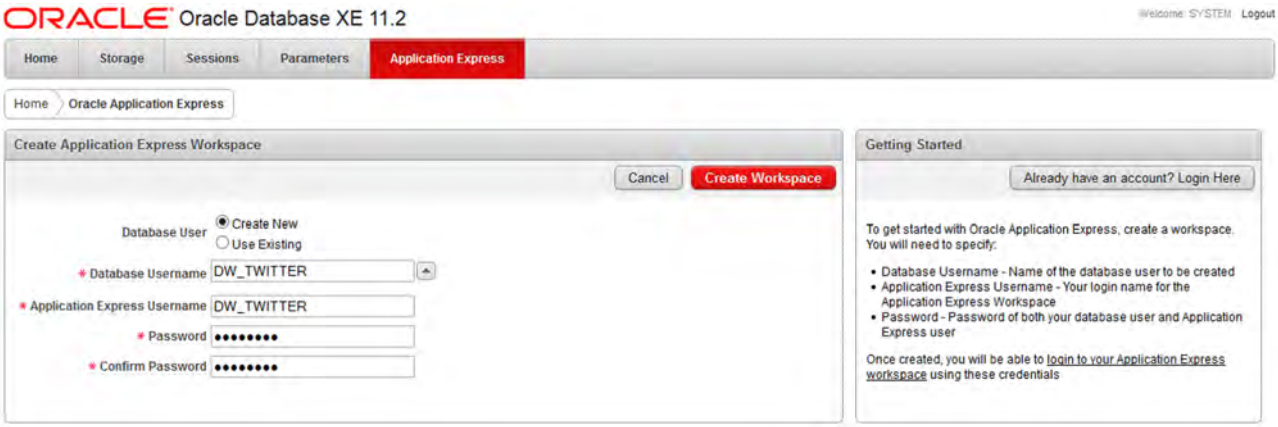
Existe un tutorial detallado con imágenes de la instalación en este otro enlace.

En primer lugar se creará un nuevo entorno de trabajo o *workspace* y un usuario nuevo (en caso de no tener ninguno). A continuación se creará cada una de las tablas definidas en el diseño conceptual.

4.3.1. Creación del entorno de trabajo

Antes de empezar con la creación de las tablas del Datawarehouse tenemos que crear un nuevo usuario y un workspace. Tanto a nuestro workspace como al usuario le daremos el nombre de DW_TWITTER.

Figura 8. Creación de usuario y Workspace en Oracle Application Express



Una vez creado el workspace y el nuevo usuario volveremos a la pantalla inicial donde veremos el siguiente mensaje:

```
Successfully created workspace DW_TWITTER. To begin
click_here to login.
```

Seremos redirigidos a la página de login donde deberemos introducir el nombre de usuario, el workspace y la contraseña.

Figura 9. Conexión a Oracle Application Express

4.3.2. Creación de las tablas del almacén de datos

Una vez conectados como el usuario DW_HOS en su workspace correspondiente, se procede a la creación de las tablas necesarias mediante SQL Workshop > Object Browser.

Para mejorar la visibilidad de las tablas con las que trabajamos se han borrado las tablas por defecto que crea el sistema.

Borrado de tablas

Para borrar cada tabla se debe seleccionar la tabla y seleccionar la opción "drop" en la parte superior del nuevo menú que aparece para la tabla.

El proceso de creación de tablas es el mismo para todas las que vamos a crear:

- Se pulsa el botón *create* situado en la parte superior derecha de la pantalla.
- Se completa el nombre de la tabla.
- Se añade la clave primaria y los diferentes atributos de la dimensión y sus características.
- Se añade la secuencia asociada a la clave primaria.
- Se añaden las claves foráneas.

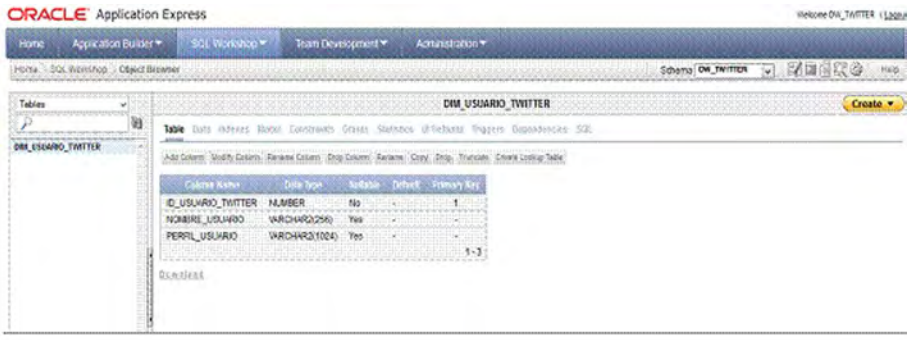
Primero se crearán las tablas de dimensión y por último las tablas de hechos.

1) Dimensión usuario: *dim_usuario_twitter*

Esta dimensión tiene dos campos:

- *id_usuario_twitter*: clave primaria, *number*.
- *nombre_usuario*: guardará el nombre del usuario, *varchar2(256)*.
- *perfil_usuario*: guardará el enlace a la cuenta del usuario en Twitter, *varchar2(1024)*.

Figura 10. Visualización del resultado de creación de una tabla con Oracle XE



2) Dimensión fecha: *dim_fecha*

Esta dimensión tiene 7 campos:

- *id_fecha*: clave primaria, *number*.
- *Año*: año de la fecha, *number(4,0)*.
- *Mes*: mes de la fecha, *number(2,0)*.
- *Dia*: día de la fecha, *number(2,0)*.
- *Hora*: hora de la fecha, *number(2,0)*.
- *Minuto*: minuto de la fecha, *number(2,0)*.
- *Segundo*: segundo de la fecha, *number(2,0)*.

- *Fecha: año/mes/día :Hora::minuto::segundo*: Contiene el valor original de la fecha, *varchar(16)*.

3) Tabla de hechos de Twitter: *h_tweet*

La tabla de hechos para almacenar los tuits está compuesta de:

- *id_tweet*: clave primaria, *number*.
- *Num_mensajes*: número de mensajes escritos por un usuarios en un día.
- *Id_fecha*: fecha en que fue escrito el tuit. Clave foránea a la tabla *dim_fecha*.
- *Id_semana*: fecha del lunes de la semana en que fue escrito el tuit. Clave foránea a la tabla *dim_Fecha*.
- *Id_usuario*: identificador del usuario que escribió el tuit. Clave foránea a la tabla *dim_usuario*.

4) Tabla de hechos de búsquedas en Google: *h_google*

El diseño físico para almacenar las búsquedas de Google solo contiene la tabla de hecho *h_google*, que contiene un atributo y una clave foránea a la dimensión fecha que hemos definido anteriormente:

- *Id_google*: clave primaria, *number*.
- *Num_busquedas*: número de búsquedas realizadas en Google en una semana, *number*.
- *Id_semana*: identificador de la semana. Clave foránea a la tabla *dim_fecha*.

5) Tabla de hechos de hospitalizaciones: *h_hosp*

La tabla de hecho está compuesta por:

- *id_hosp*: clave primaria, *number*.
- *episodio*: el número de historia clínica del paciente, *varchar2(14)*.
- *nhc*: el identificador de historia clínica del paciente, *varchar2(12)*.
- *diagnostico*: descripción de la patología, *varchar2(512)*.
- *dias_hospitalizacion*: días totales de hospitalización, *number*.
- *id_fecha_ingreso*: identificador de la fecha de ingreso del paciente. Clave foránea a la tabla *dim_fecha*.

5. Diseño e implementación de la carga de datos en el almacén de datos

Una vez creado el almacén de datos es el momento de centrarnos en la creación de los procesos ETL para cargar sus datos. Como ya sabemos, estos procesos consisten en la extracción, transformación y carga de los datos. En definitiva, lo que se persigue es estructurar, integrar y acomodar los datos de las fuentes de origen en el almacén de datos.

En nuestro caso particular tenemos tres fuentes de origen en ficheros Excel, y una fuente de destino, Oracle, la base de datos donde se aloja nuestro almacén de datos.

Para introducir los datos a partir de las diferentes fuentes de origen, primero identificaremos los diferentes procesos de ETL necesarios, indicando si la carga de datos se realiza una única vez o se deben realizar cargas incrementales. Seguidamente, definiremos todos los pasos que se realizarán en cada proceso ETL para cada una de las dimensiones o tablas de hechos y, para finalizar, implementaremos estos pasos de extracción y carga de datos con Oracle Application Express y Visual Studio.

5.1. Identificación de procesos ETL necesarios

Los procesos ETL deben conceptualizarse como manipulaciones de flujos de datos. Estos procesos deben diseñarse teniendo en cuenta diversos factores como:

- Cómo debe cargarse de forma lógica la información, es decir, qué debe cargarse primero y qué después.
- La ventana de tiempo disponible, hecho que puede condicionar lo que debemos cargar.
- Tipo de carga: inicial o incremental.

En nuestro caso tenemos cuatro procesos de carga diferentes:

- Carga inicial del conjunto de datos de tuits, búsquedas en Google y hospitalizaciones.
- Proceso de carga diario del conjunto de tuits del día anterior.
- Proceso de carga semanal del número de búsquedas sobre la gripe realizadas en Google.
- Proceso de carga en periodos de tiempo más elevados del conjunto de datos de las hospitalizaciones.

El primer paso en la creación de los procesos ETL es comprender qué procesos son necesarios y en qué orden deben realizarse, teniendo en cuenta que:

- Es necesario identificar si los datos deben cargarse en un área intermedia. En nuestro caso, no es necesario, para cada proceso de carga de datos (Twitter, Google y hospitalizaciones), los datos serán cargados en tablas de dimensiones y hechos diferentes, a excepción de la tabla fecha.

- Los valores que se insertarán en las tablas de dimensiones son valores no fijos que se extraerán, transformarán y cargarán mediante procesos ETL. Estas dimensiones son *usuario_twitter* y *fecha*.

- Por último, las tablas de hechos tuit, Google y hospitalización también se cargarán mediante un proceso ETL.

Para respetar la integridad referencial del diseño físico del almacén de datos, las dimensiones se cargarán antes que las tablas de hecho.

5.2. Diseño de los procesos ETL

En este apartado vamos a describir la funcionalidad a alto nivel de cada uno de los procesos ETL identificados. Tener claro a alto nivel qué debe hacer el ETL ayuda a posteriori en el proceso de diseño usando una herramienta específica.

1) Carga inicial de la dimensión *usuario Twitter*

El proceso ETL para la carga de datos en la dimensión *usuario Twitter* consistirá en:

- Extraer los valores del fichero Excel.
- Omitir los campos innecesarios.
- Ordenar los valores y eliminar los valores duplicados.
- Cargar los datos en la tabla correspondiente.

2) Carga inicial de la dimensión *fecha*

Además de los pasos descritos para la dimensión *usuario Twitter*, se añadirá un nuevo paso para calcular los datos derivados de la tabla de dimensión. Es decir, a partir de cada fecha, se calculará (y cargará en la base de datos) el mes, el año, la semana, el día, la hora, el minuto y el segundo de la misma. El proceso ETL consistirá en:

- Extraer los valores del fichero Excel.
- Omitir los campos innecesarios.
- Ordenar los valores y eliminar los valores duplicados.
- Cálculo de atributos derivados.

- Cargar los datos en la tabla correspondiente.

3) Carga inicial de las tablas de hechos

Todas las tablas de hechos tendrán las mismas subtareas para la carga de datos:

- Extraer valores.
- Omitir campos innecesarios.
- Recuperar *id* para cada una de las dimensiones.
- Cargar tabla de hecho.

4) Cargas de datos incrementales

Además de la carga de datos inicial, debemos definir los procesos de carga de datos incrementales de las diferentes fuentes de datos.

Para el fichero de datos de Twitter, se cargarán datos de forma incremental en la dimensión *usuario Twitter* y en tabla de hechos de tuits.

El proceso incremental consistirá en recuperar los datos de Twitter tal y como se ha explicado en la primera parte del caso y se realizarán los mismos procesos descritos en la carga de datos iniciales:

- Extraer valores.
- Omitir campos innecesarios.
- Cálculo de atributos derivados (solo para la dimensión *fecha*).
- Recuperar *id* para cada una de las dimensiones (solo en el caso de la tabla de hechos de tuits).
- Cargar tabla de dimensión o hechos.

Para realizar el proceso de carga incremental de los datos de Google Flu Trends, se deberá añadir un primer paso que consistirá en la recuperación de información de Google. Google devuelve los datos sobre las búsquedas de la gripe en formato CSV, así que el paso de recuperación de la información consistirá en:

- Descargarse los datos de las búsquedas realizadas sobre gripe.
- Verificar que el fichero se ha actualizado desde la última inserción.
- Recuperar el último valor correspondiente a la región de Cataluña.
- Insertar los datos en la dimensión fecha y en la tabla de hechos tal y como se ha descrito anteriormente.

El proceso de carga incremental para los datos de hospitalización no ha sido creado debido a que no está clara la posibilidad de obtener estos datos periódicamente ni, de ser posible, la periodicidad en que se podrían conseguir. No obstante, la carga incremental de hospitalizaciones se haría de forma parecida a la de Twitter.

5.3. Implementación de los procesos de carga de datos

Para realizar la implementación de los procesos ETL vamos a utilizar Microsoft Visual Studio e Integration Services.

El esquema que seguiremos en la implementación será:

- Conocer el entorno de trabajo y crear un nuevo proyecto.
- Conexión a las fuentes de datos de origen y de salida.
- Implementación de cada uno de los procesos ETL descritos en el apartado anterior.

5.3.1. Creación de un nuevo proyecto

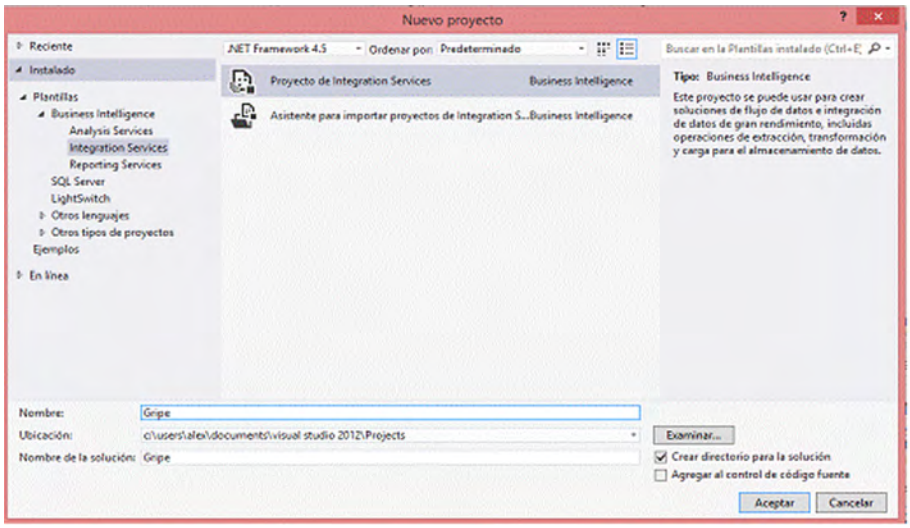
Antes de empezar a trabajar, y de cara a evitar problemas de acceso a la base de datos que hemos creado con Oracle Apex desde Microsoft Visual Studio, es muy importante ejecutar Microsoft Visual Studio con permisos de administrador.

Modo administrador

Para ejecutar Microsoft Visual Studio como administrador hacemos clic sobre el icono del programa con el botón derecho del ratón y seleccionamos la opción de ejecutar como administrador.

Al ejecutar Microsoft Visual Studio si iniciamos un nuevo proyecto podremos cargar un proyecto ya existente o crear un nuevo proyecto. Para la implementación de los procesos ETL crearemos un nuevo proyecto de Integration Services.

Figura 11. Inicio de un nuevo proyecto de Integration Services



Una vez creado el proyecto es importante familiarizarse con las diferentes zonas de trabajo que tiene el entorno de Integration Services:

- Cuadro de herramientas: tareas ETL disponibles (en la parte izquierda de la pantalla).

- Explorador de soluciones: estructura del proyecto (en la parte superior derecha de la pantalla).

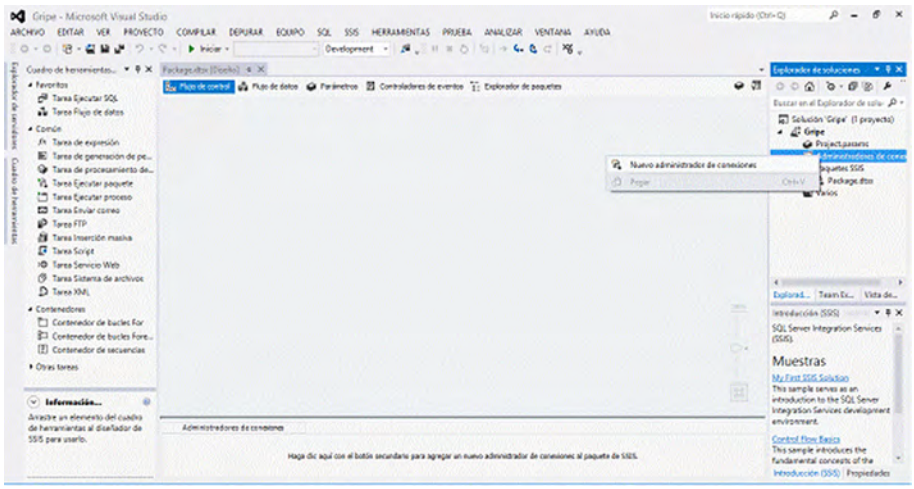
- Explorador de elemento: que muestra las propiedades de este (en la parte inferior derecha de la pantalla).

- Área de trabajo: en la que se definen los paquetes de ETL (en la parte central de la pantalla). Tenemos diferentes opciones: flujo de control, flujo de datos, parámetros, controladores de eventos y explorador de paquetes.

5.3.2. Conexión a las fuentes de datos de origen y destino

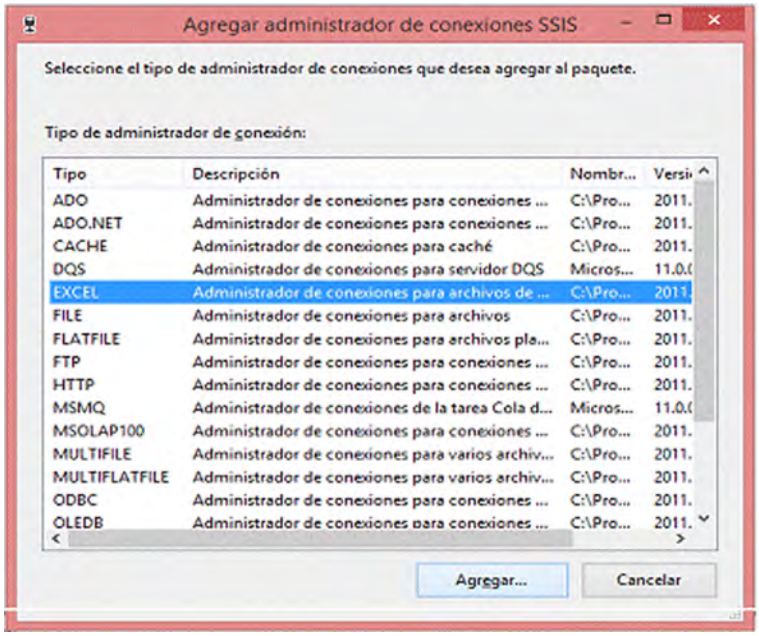
Los procesos ETL que vamos a implementar trabajarán con alguna de las tres fuentes de origen (Excel) y con una fuente de destino (Oracle). Para que estas fuentes de origen y destino estén disponibles para todo el proyecto, en el explorador de soluciones > *Administración de conexiones* añadimos todas ellas seleccionando la opción "Administrador de conexiones" con el botón derecho del ratón.

Figura 12. Creación de un nuevo administrador de conexiones



Para crear las conexiones de origen con los ficheros Excel añadimos para cada uno de ellos una nueva conexión y seleccionamos la conexión de tipo Excel. Tendremos que indicar la ruta del fichero cada vez.

Figura 13. Agregación de administrador de conexiones Excel



La creación de la conexión con la fuente de datos de destino Oracle es mucho más sencilla. De la misma manera que para la fuente de origen, tendremos que seleccionar el explorador de soluciones > *Administración de conexiones*, seleccionar la opción *Administrador de conexiones* con el botón derecho del ratón, crear nueva conexión de datos y esta vez escogemos la opción *ADO.NET*.

Una vez seleccionada esta opción, nos aparecerá un menú donde tendremos que seleccionar *Proveedor.NET\OracleClient Data Provider* y rellenar los datos del usuario y contraseña que hemos indicado en la base de datos de Oracle XE al crear el diseño físico del modelo de datos.

5.3.3. Implementación de los procesos ETL

Tal y como se ha comentado anteriormente, la mayoría de los procesos ETL realizan las siguientes subtareas:

- Extraer los valores del fichero origen.
- Omitir los campos innecesarios.
- Ordenar los valores y eliminar los valores duplicados.
- Cargar los datos en la tabla correspondiente.

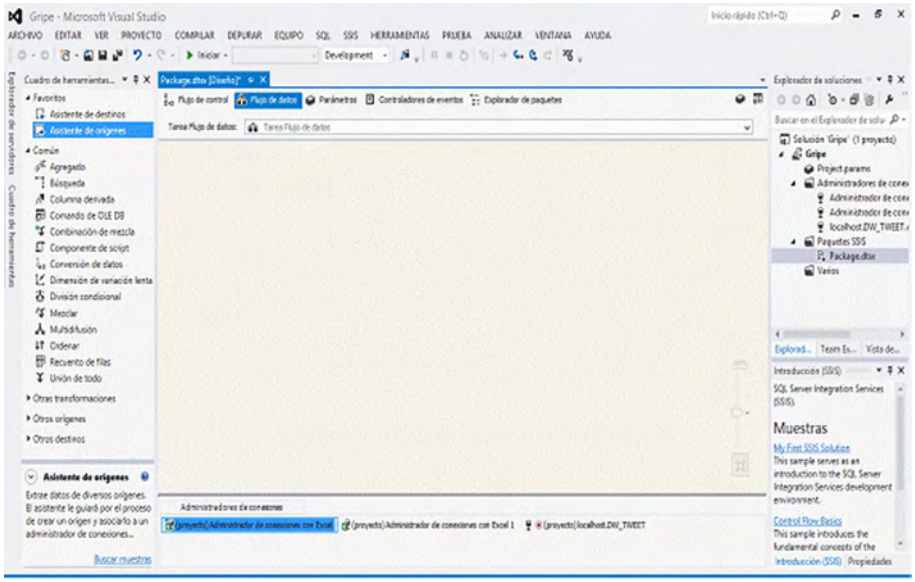
A continuación se muestra cómo implementar los procesos ETL de carga inicial definidos en Visual Studio. Se empezará con la implementación de los procesos que cargan datos en las dimensiones y se finalizará con el proceso que carga los datos en las tablas de hechos.

1) Carga inicial de la dimensión *usuario Twitter*

Antes de empezar a realizar las subtareas, debemos crear un nuevo *package* para agrupar todas las subtareas del proceso ETL. Para ello, en el menú superior seleccionamos la pestaña *proyecto* y la subpestaña *nuevo paquete de SSIS* y renombramos al paquete creado a dimensión *usuario Twitter*.

Posteriormente crearemos el paso para extraer los valores del fichero origen. Para ello creamos una tarea de flujo de datos, seleccionando la segunda opción justo encima del contenedor central del área de trabajo.

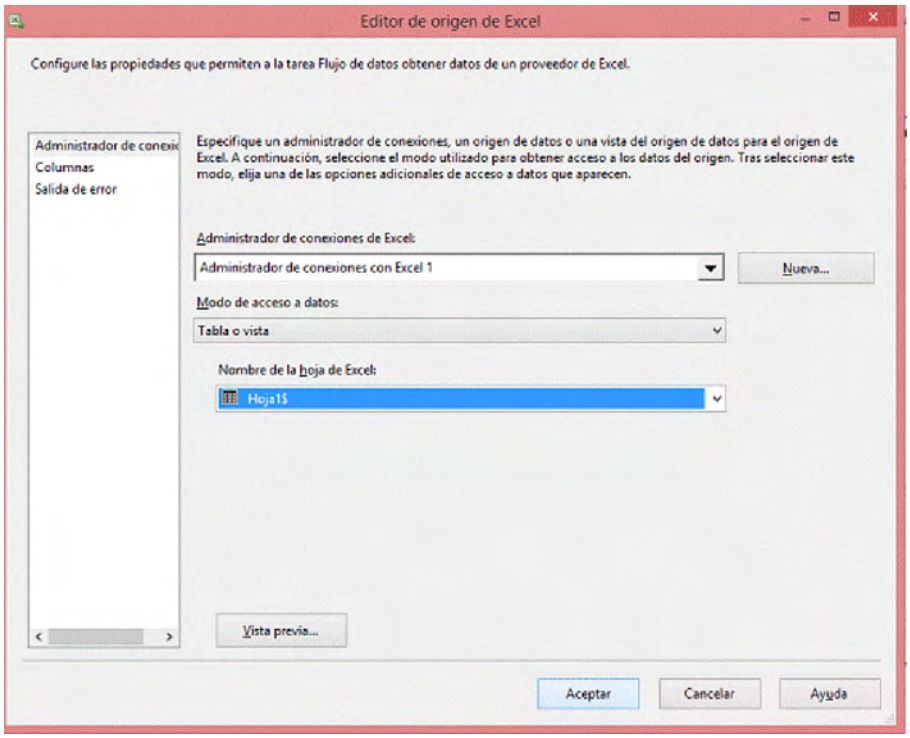
Figura 14. Creación de un flujo de datos



Una vez seleccionada la tarea de flujo de datos, seleccionamos el asistente de orígenes que se encuentra en la parte superior izquierda. Se abrirá una nueva ventana de diálogo en la que seleccionaremos el flujo de datos de la hoja de Excel que contiene los valores de Twitter.

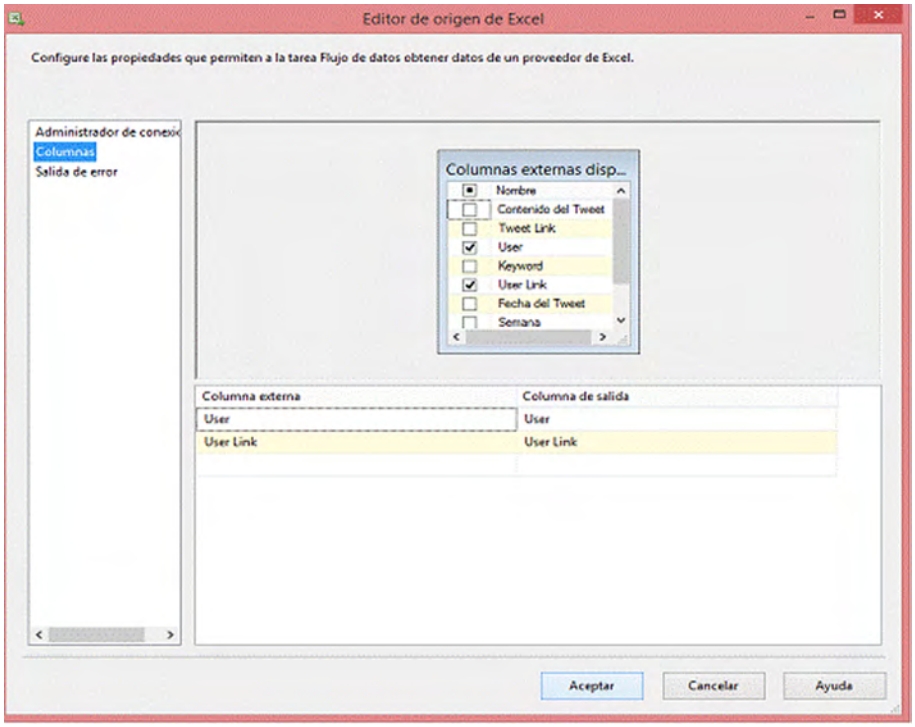
En el área de trabajo se creará un objeto. Si el fichero de origen Excel tiene más de una hoja, como será el caso del fichero de las hospitalizaciones, tendremos que seleccionar el nombre de la hoja que se utilizará.

Figura 15. Configuración de flujo de datos Excel



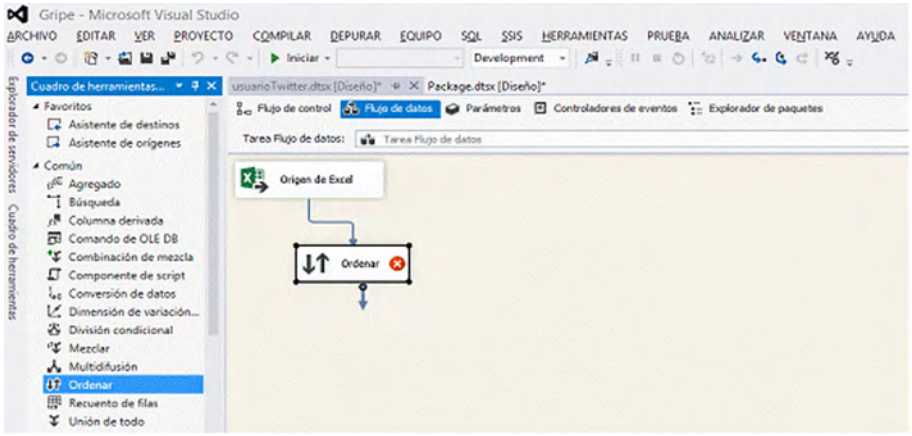
Para omitir los datos innecesarios del fichero de origen, en el mismo diálogo en el que hemos seleccionado el nombre de la hoja de Excel en el paso anterior, seleccionamos la pestaña *Columnas*. Una vez allí, seleccionamos las columnas que se seleccionan: *usuario* y *user link*.

Figura 16. Asignación de equivalencias entre el origen de datos y el flujo de datos



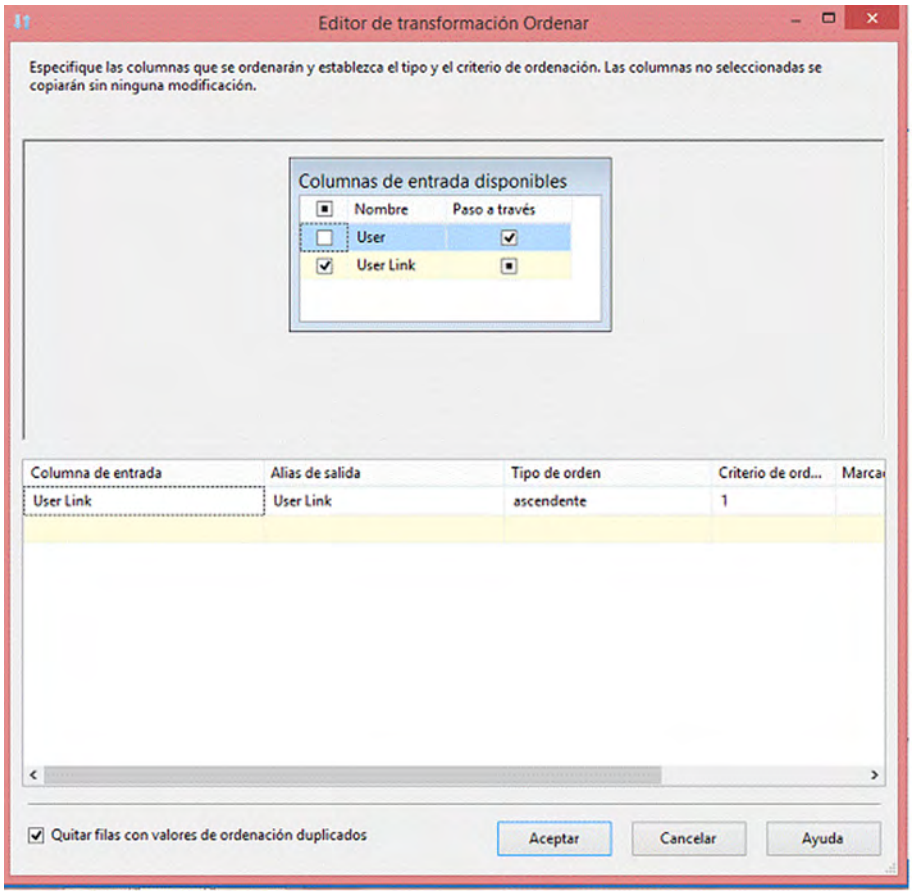
Posteriormente se deberá ordenar los datos y eliminar los duplicados. Para ello en al cuadro de herramientas del menú de la izquierda seleccionamos la opción *ordenar* y unimos el origen de datos Excel con el criterio de ordenación.

Figura 17. Adición de un paso de ordenación al flujo de datos



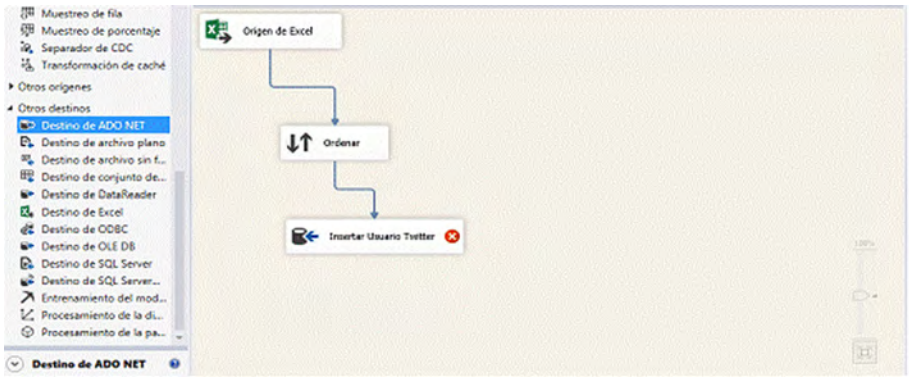
Hacemos clic en el icono de *ordenar* del área de trabajo y escogemos el atributo *user_link*. Esto indica que queremos ordenar los datos por el campo *user_link*. Seleccionamos la opción de *Quitar filas con valores de ordenación duplicados* para eliminar los valores duplicados.

Figura 18. Selección de las columnas a ordenar y eliminación de datos repetidos



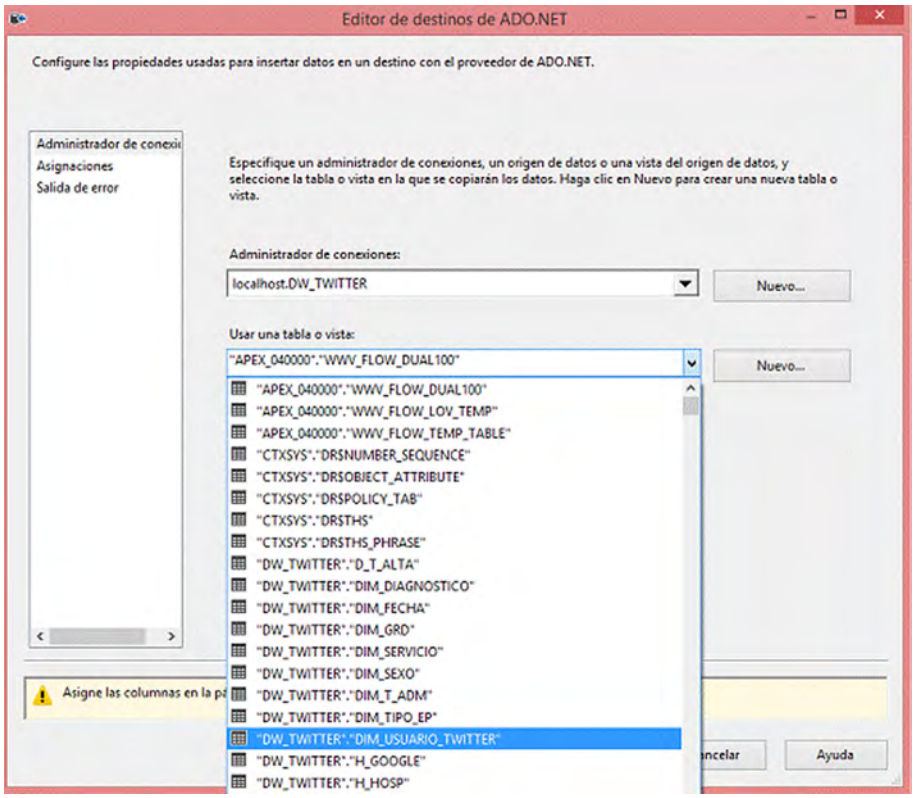
Para añadir los datos a la tabla de dimensión deberemos añadir un paso de inserción, seleccionando la opción *Otros destinos* > *Destino de ADO NET*.

Figura 19. Adición en el flujo de datos de un paso de inserción de datos en el *datawarehouse*



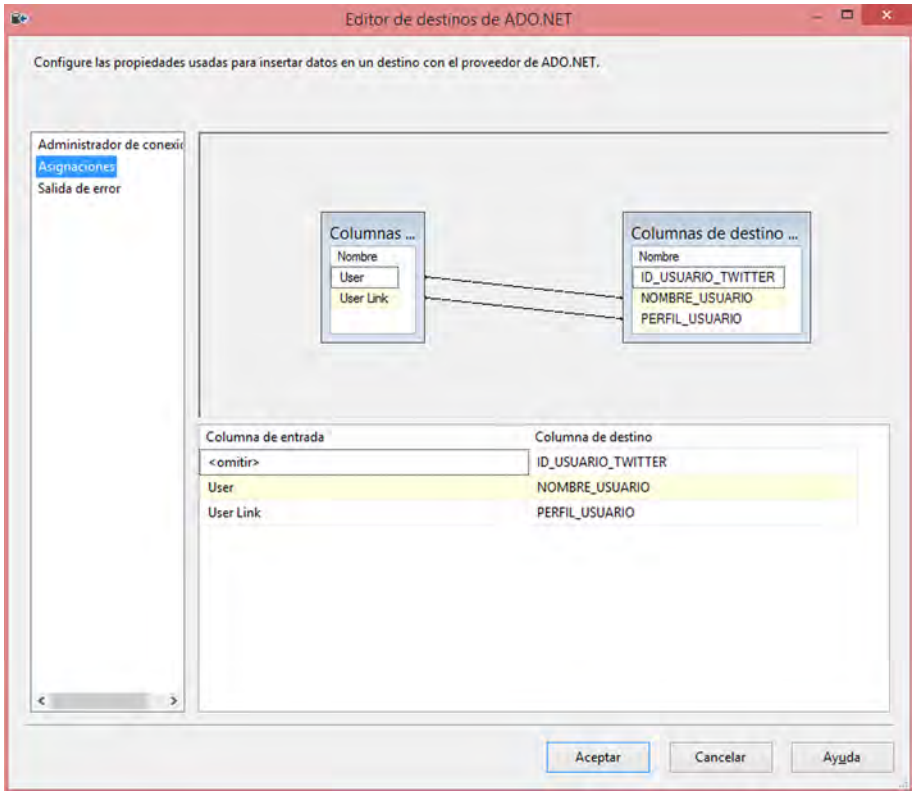
Hacemos clic en el objeto que se acaba de crear en el área de trabajo y lo configuramos correctamente para que apunte a la dimensión correspondiente (*DIM_USUARIO_TWITER*).

Figura 20. Selección de la tabla del almacén de datos en el *datawarehouse*



Posteriormente, para indicar dónde se debe guardar cada campo del fichero origen, asignamos los campos de entrada de origen con las columnas de la tabla de destino.

Figura 21. Asignación de los campos de entrada de origen con las columnas de la tabla de destino



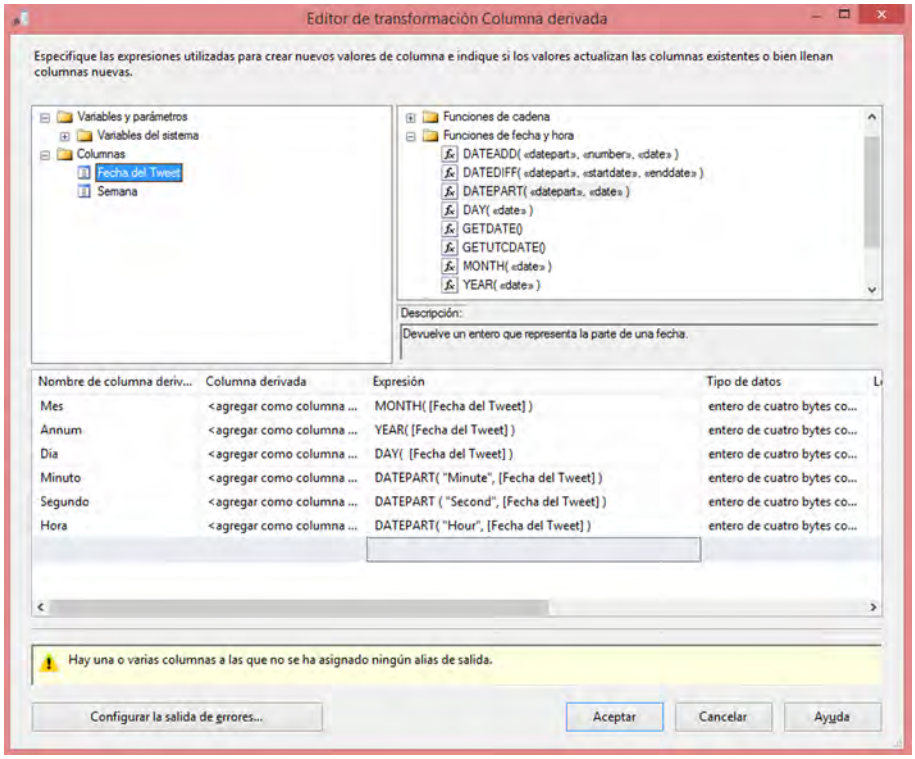
2) Carga inicial de la dimensión *fecha*

Para crear el proceso ETL de la dimensión *fecha* empezamos creando las mismas subtareas que hemos empleado en el proceso ETL dimensión *usuario Twitter*:

- Creamos el origen de datos.
- Seleccionamos el atributo *fecha* del tuit.
- Añadimos un paso de ordenación y quitamos los elementos duplicados.

Una vez llegados a este punto, a partir de la fecha calculamos algunos valores derivados para permitir consultar la tabla de hechos con diferentes niveles de agregación. Para ello añadimos un paso de tipo *Columna derivada* para generar columnas derivadas y almacenar en la dimensión *fecha* los campos *año*, *mes*, *día*, *hora*, *minuto* y *segundo* a partir del valor de la fecha del tuit. En la figura siguiente podemos ver los campos calculados y las fórmulas usadas para calcular sus valores.

Figura 22. Creación de un paso de columna derivada para descomponer los campos de las fechas

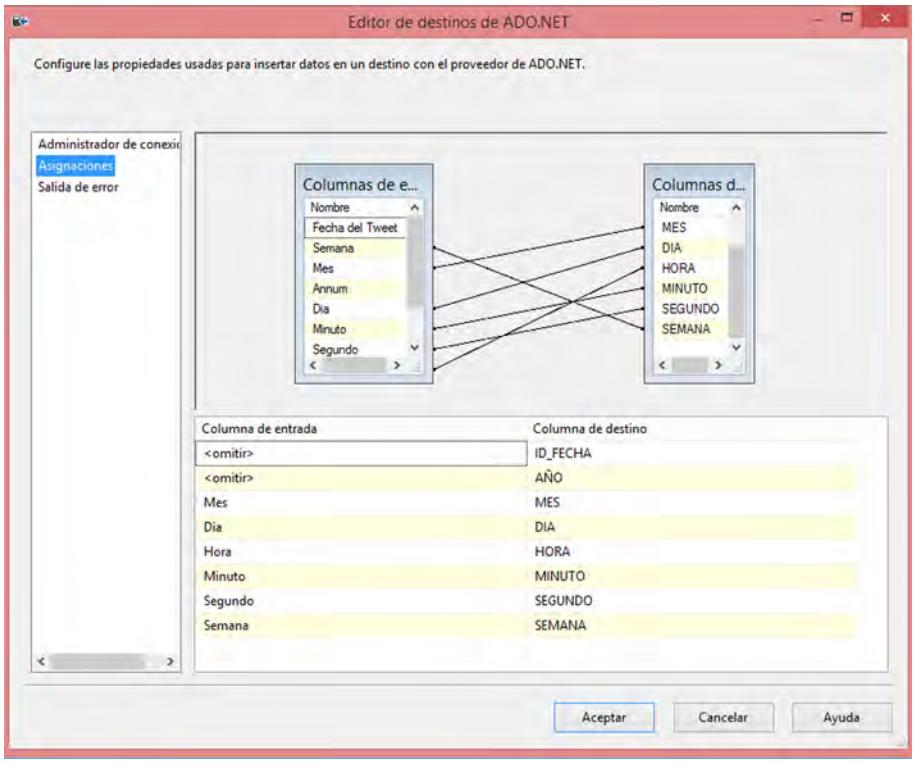


La última subtarea del proceso ETL consiste en añadir el paso de inserción, seleccionando *Otros destinos* > *Destino de ADO.NET*, tal y como ya hemos hecho en la dimensión *usuario Twitter*:

- Seleccionamos *Oracle* como fuente de destino y escogemos la tabla dimensión *fecha* en la base de datos de Oracle.
- Asignamos cada columna del fichero de origen con su columna correspondiente en la dimensión *fecha* del almacén de datos.

Una vez creados los procesos ETL podemos ejecutarlos para cargar los datos o planificar su ejecución periódicamente.

Figura 23. Asignación de columnas de tipo fecha entre el origen y el destino



3) Carga inicial de las tablas de hechos

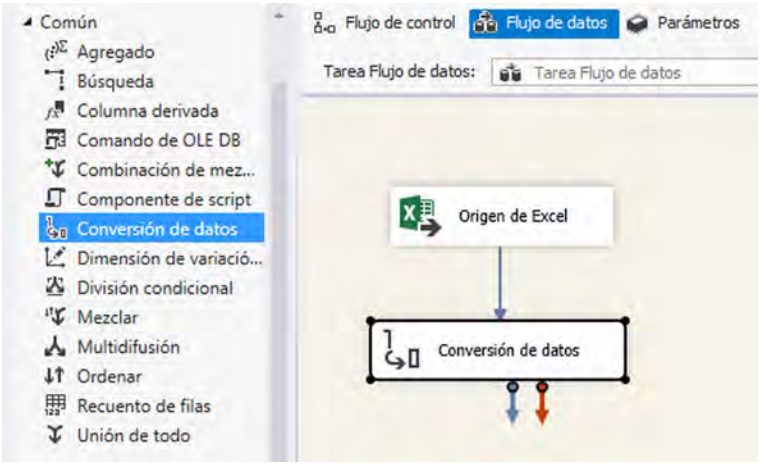
Para implementar el proceso ETL de las tablas de hechos realizamos tareas anteriormente descritas:

- Extraer valores.
- Omitir campos innecesarios.
- Recuperar el identificador para cada una de las dimensiones.
- Cargar los datos en las tablas de hecho.

Los dos primeros pasos son idénticos a los indicados en los procesos ETL de implementación de las dimensiones. Vamos a centrarnos en los pasos que difieren de los presentados anteriormente.

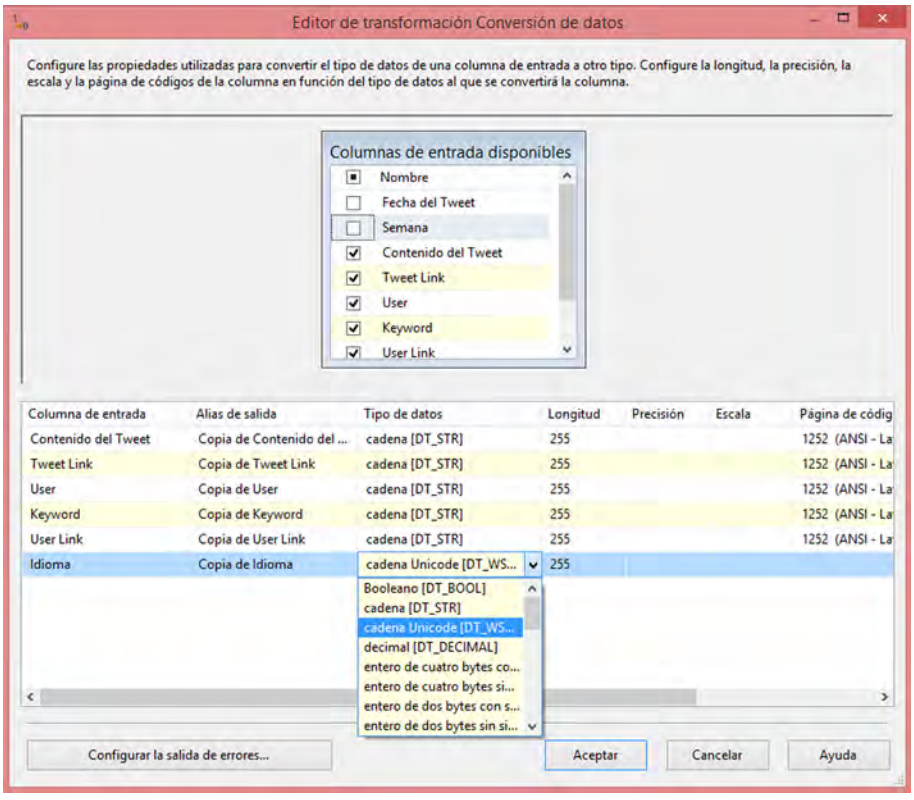
Los datos en la fuente de origen de Excel están codificados en Unicode. Para evitar problemas de incompatibilidad deberemos crear un paso intermedio de *Conversión de datos* en el proceso ETL.

Figura 24. Adición de un paso de conversión de datos



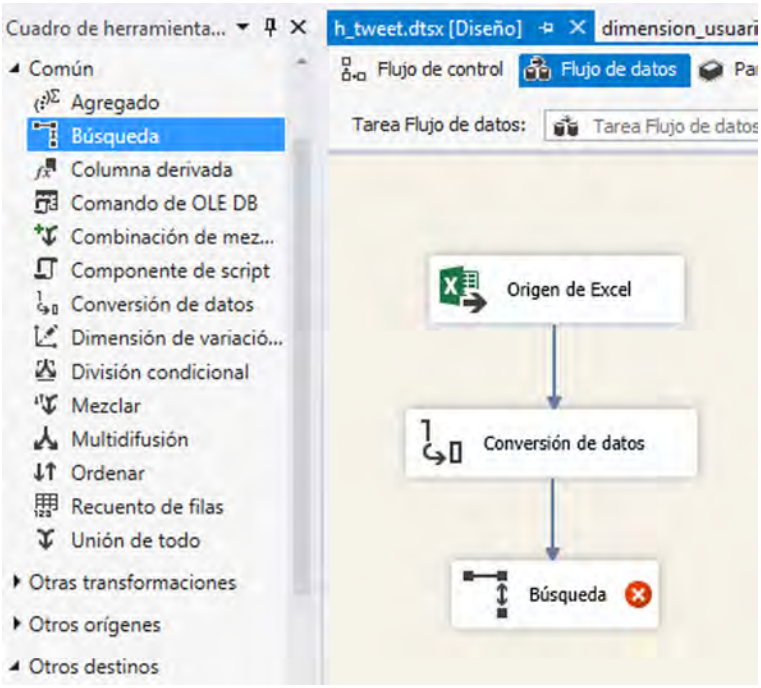
En la conversión de datos seleccionamos todos los campos del fichero de tipo *DT_WSTR* (cadena de texto en Unicode) e indicamos que se deberá convertir en una versión sin Unicode (no unicode).

Figura 25. Selección de las columnas para la conversión de datos



Para poder dar de alta un registro en la tabla de hechos será necesario tener los valores de las claves foráneas que contiene. Para ello, se deberá obtener, para cada clave foránea, el identificador del registro con el que el hecho está relacionado. Para ello, utilizaremos un paso de *búsqueda* en el proceso ETL. Este paso encontrará, en la tabla de la dimensión correspondiente, el registro asociado a los datos que hay en el fichero de origen.

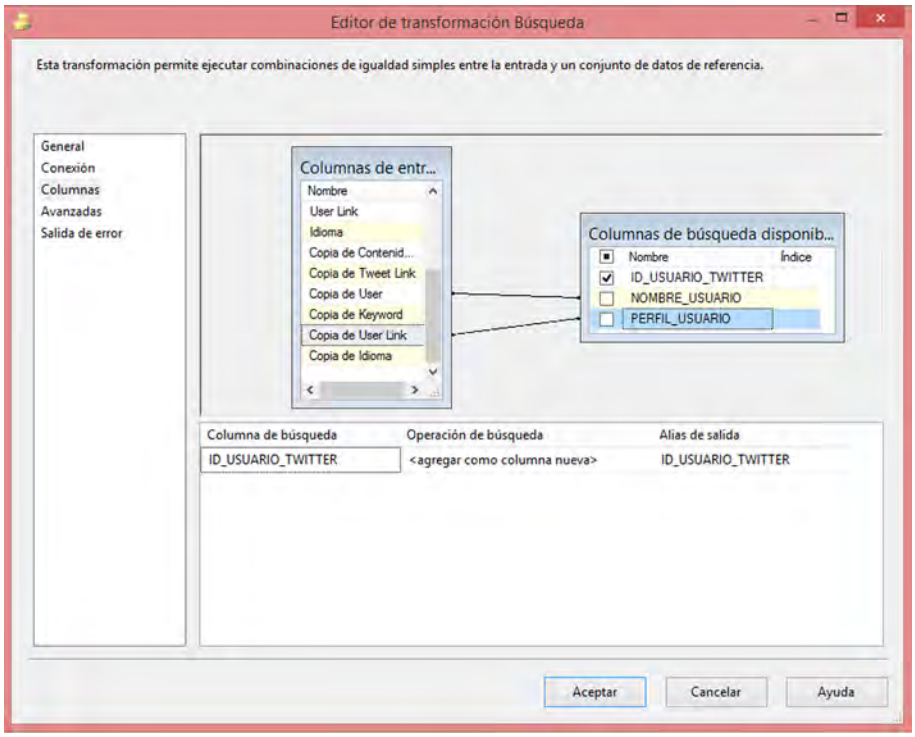
Figura 26. Adición de un paso para la búsqueda de las claves foráneas



Cuando seleccionemos el cuadro de búsqueda por primera vez, el sistema pedirá que se cree la conexión a la base de datos. Podemos crear tantas conexiones como creamos y la conexión que se realizará para este caso práctico es la misma conexión *ADO.NET* que hemos creado en repetidas ocasiones anteriormente.

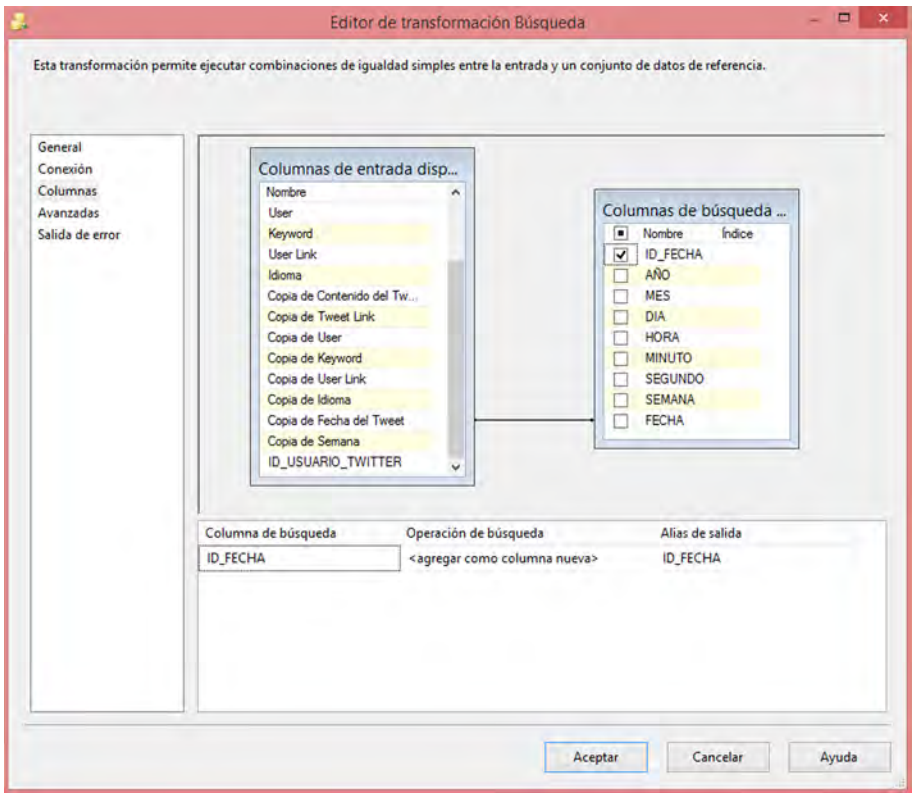
A modo de ejemplo, en la siguiente figura se muestra cómo realizar la búsqueda de la clave primaria de un usuario de Twitter a partir de los datos de entrada.

Figura 27. Selección de las claves primarias y foráneas de las tablas



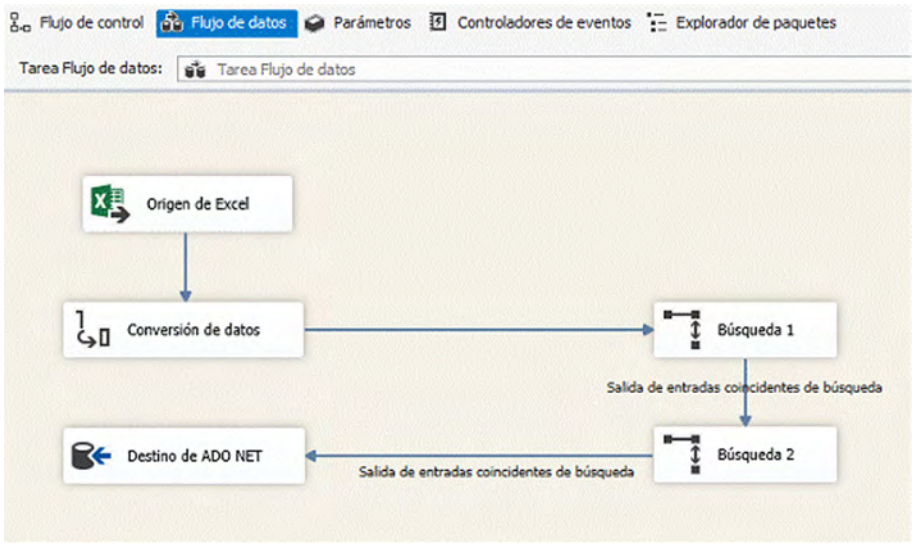
Una vez relacionadas la tabla de tuits con la tabla de la dimensión *usuario Twitter*, debemos relacionar también la tabla de hechos de tuits con la tabla de la dimensión *fecha*. Para ello, debemos crear un paso de *búsqueda* nuevo para relacionar el campo de origen *Copia de fecha del tuit* con el atributo *fecha* de la dimensión *usuario_tweet*.

Figura 28. Selección de las claves primarias y foráneas para el capo fecha en las diferentes tablas



La última subtarea que se realizará para finalizar el proceso ETL será insertar los datos en la tabla de hechos de tuits. Para ello, añadimos un paso de inserción *ADO.NET*, tal y como hemos visto antes. El proceso ETL resultante quedaría de la siguiente manera:

Figura 29. Creación de un paso de inserción delos datos en el *datawarehouse*



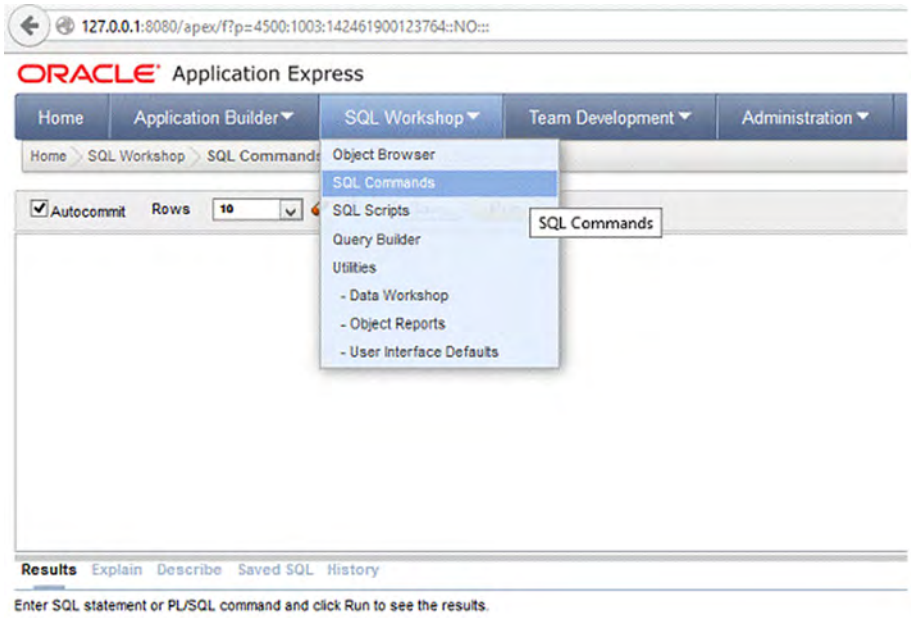
6. Implementación del cuadro de mando

En este apartado veremos cómo implementar el cuadro de mando diseñado. No obstante, antes mostraremos las consultas que nos permitirán calcular cada uno de los indicadores integrados en el cuadro de mando.

6.1. Cálculo de los indicadores del cuadro de mando

Para probar las consultas que permiten calcular los valores de los indicadores del cuadro de mando, ejecutamos la aplicación de Oracle Application Express que hemos utilizado para implementar el almacén de datos y seleccionamos la pestaña *SQL Commands*.

Figura 30. Acceso a la consola de comados SQl en Oracle Application Express



Una vez allí podemos verificar el correcto funcionamiento de las consultas para obtener los datos requeridos.

A continuación podemos ver las consultas que permiten calcular los datos necesarios de cada uno de los elementos del cuadro de mando diseñado.

1) Número de mensajes escritos en Twitter y de consultas realizadas en Google durante la última semana:

```
SELECT SUM(t.num_mensajes) as tweets,
      g.NUMERO_BUSQUEDAS as busquedas, f.semana
FROM H_TWEET t, H_GOOGLE g, DIM_FECHA f
WHERE t.ID_FECHA = f.ID_FECHA and g.ID_FECHA = f.ID_FECHA and
      f.FECHA > sysdate -7
```

```
GROUP BY f.semana
```

2) Probabilidad de estar en un periodo que precede a un brote de gripe:

```
SELECT CASE WHEN
   SUM(t.num_mensajes) > 75 and
   (SUM(t.num_mensajes) - SUM(t2.num_mensajes) ) / SUM(t2.num_mensajes) > 2
   THEN 'Alta'
   ELSE 'Baja'
   END as probabilidad
FROM H_TWEET t, H_TWEET t2, DIM_FECHA f1, DIM_FECHA f2
WHERE
   t.ID_FECHA = f1.ID_FECHA and t2.ID_FECHA = f2.ID_FECHA and
   f1.FECHA> sysdate -7 and
   f2.FECHA > sysdate -7*2 and f2.FECHA < sysdate -7*3
GROUP BY f1.semana, f2.semana
```

3) Número de tuits por semana escritos en los últimos 6 meses:

```
SELECT SUM(t.num_mensajes) as tweets, f.semana as semana
FROM H_TWEET t, DIM_FECHA f
WHERE t.ID_FECHA = f.ID_FECHA and f.fecha > add_months( sysdate, -6)
GROUP BY f.semana
ORDER BY f.semana
```

Las funciones *add_months* y *sysdate* permiten calcular una fecha de forma automática a partir de la fecha actual.

4) Mensajes escritos en Twitter y las búsquedas realizadas en Google sobre la gripe en el último año:

```
SELECT SUM(t.num_mensajes) as tweets, g.numero_busquedas as busquedas,
   f.semana
FROM H_TWEET t, H_GOOGLE g, DIM_FECHA f
WHERE t.ID_FECHA = f.ID_FECHA and g.ID_FECHA = f.ID_FECHA
   and f.FECHA > add_months( sysdate, -12)
GROUP BY f.semana
ORDER BY f.semana
```

5) Mensajes semanales en Twitter sobre gripe de dos años atrás:

```
SELECT SUM(t.num_mensajes) as tweets, f.semana as semana
FROM H_TWEET t, DIM_FECHA f
WHERE t.ID_FECHA = f.ID_FECHA and f.fecha > add_months( sysdate, -24) and
   f.fecha < add_months( sysdate, -12)
GROUP BY f.semana
```

```
ORDER BY f.semana
```

6) Variación porcentual semanal del número de mensajes escritos en Twitter (la variación porcentual en intervalos de dos semanas se calcularía igual pero restando 14 días a la fecha en vez de 7):

```
SELECT ROUND ( (( SUM(t.num_mensajes) &#8211; SUM(t2.num_mensajes) )
      / SUM(t2.num_mensajes) )*100, 2 ), f1.semana
FROM H_TWEET t, H_TWEET t2, DIM_FECHA f1, DIM_FECHA f2
WHERE
   t.ID_FECHA = f1.ID_FECHA and t2.ID_FECHA = f2.ID_FECHA and
   f2.FECHA = to_date (f1.fecha)-7
GROUP BY f1.semana
ORDER BY f1.semana
```

7) Variación porcentual semanal del número de mensajes escritos en Google (la variación porcentual en intervalos de dos semanas se calcularía igual pero restando 14 días a la fecha en vez de 7):

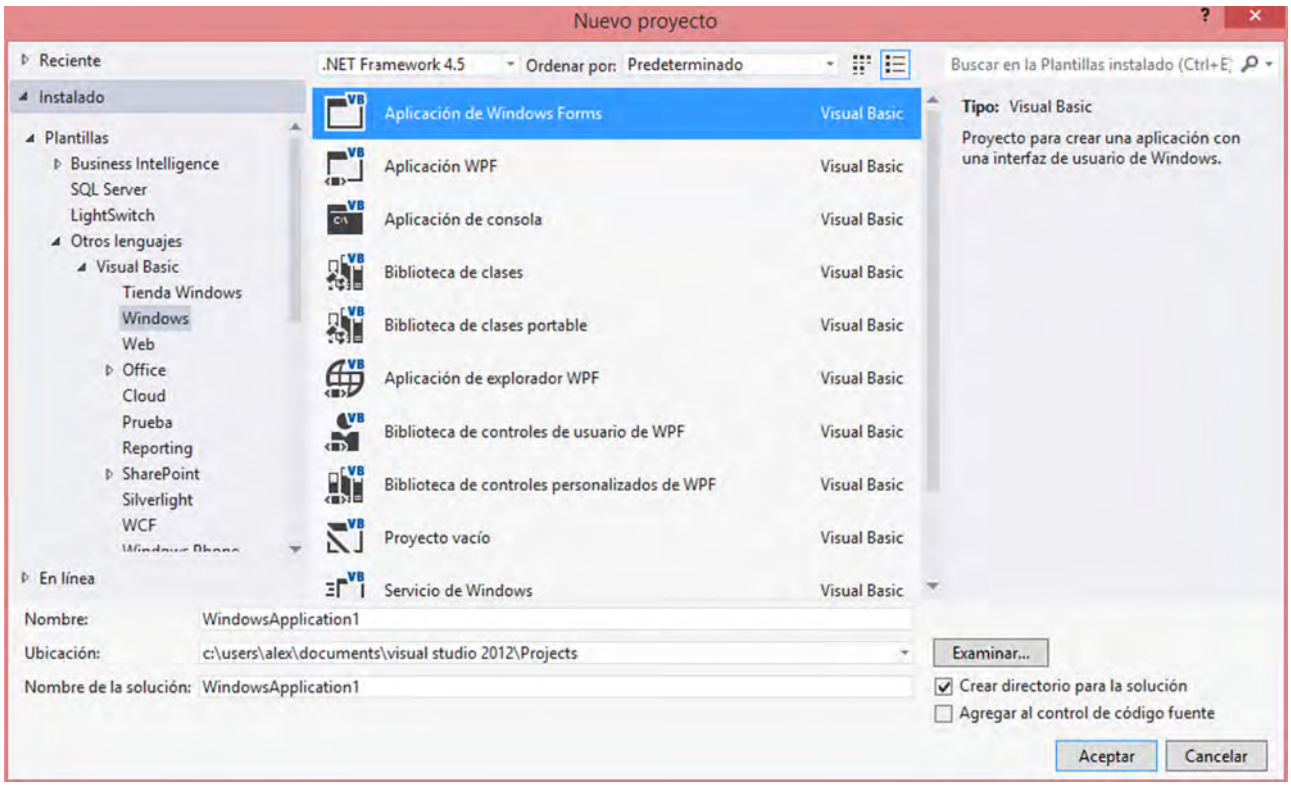
```
SELECT round( ( ( g1.numero_busquedas - g2.numero_busquedas)
      / g2.numero_busquedas) * 100 , 2), f1.semana
FROM h_google g1, h_google g2, dim_fecha f, dim_fecha f2
WHERE
   g1.ID_FECHA = f.id_fecha and g2.ID_FECHA = f2.id_fecha and
   f2.fecha = to_date(f.fecha)-7
GROUP BY f1.semana
ORDER BY f1.semana
```

6.2. Implementación gráfica del cuadro de mando

Una vez introducidos los datos en nuestro almacén de datos y creadas las consultas que permitirán extraer la información de este, pasamos a implementar de forma gráfica el cuadro descrito en la segunda parte del caso práctico.

Para ello crearemos un nuevo proyecto en Visual Studio, escogiendo un proyecto del tipo *Aplicación de Windows Form*.

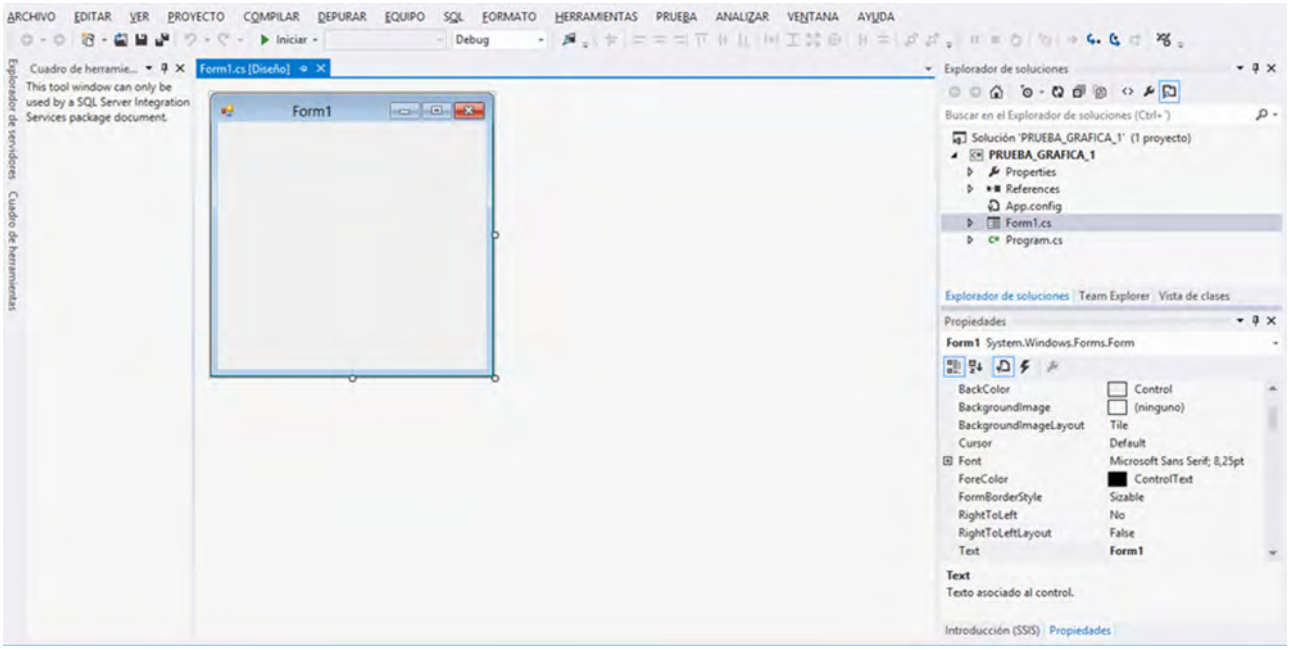
Figura 31. Creación de una Aplicación visual en Microsoft Visual Studio



Al crear el proyecto se creará un formulario de interacción. Este formulario será la pantalla principal de nuestro cuadro de mandos y, aun estando inicialmente vacío, acabará conteniendo:

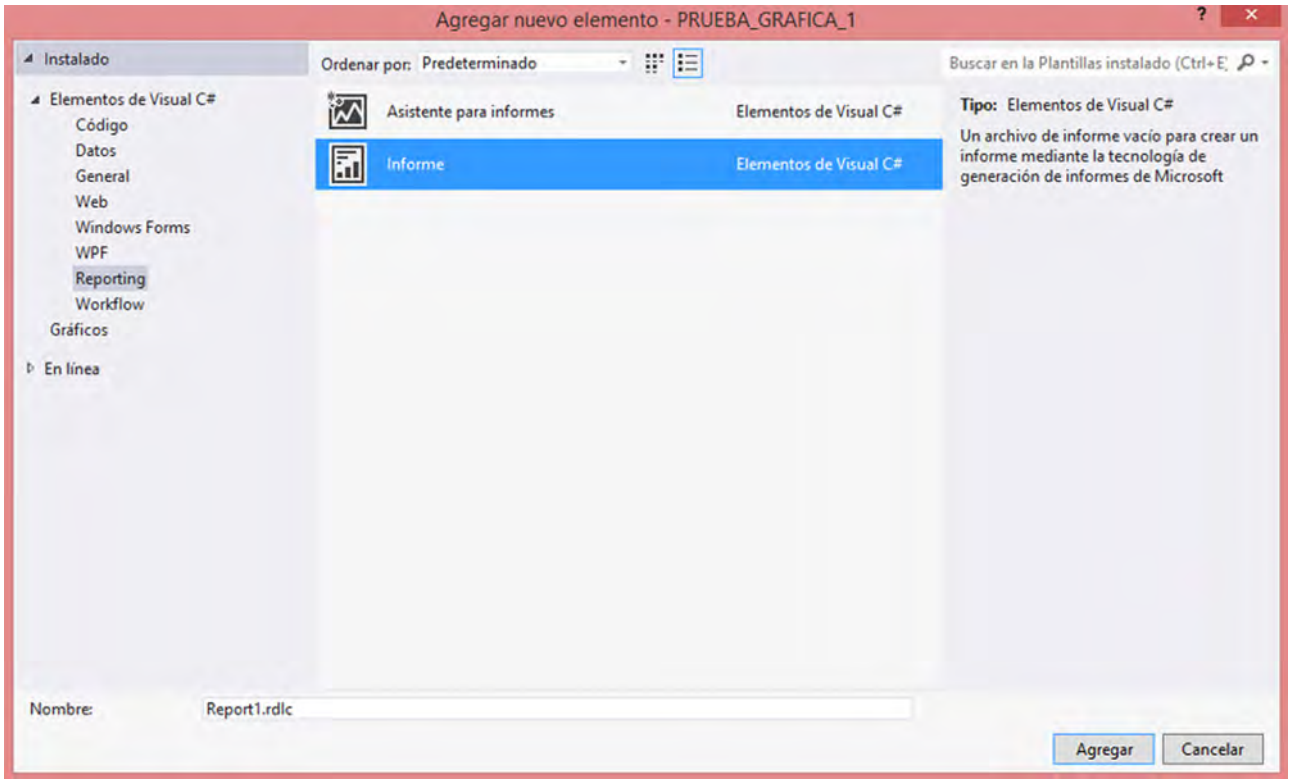
- La fecha actual.
- Un gráfico con los tuits por semana escritos sobre la gripe en los últimos 6 meses.
- Un gráfico con los mensajes escritos en Twitter y las búsquedas realizadas en Google sobre la gripe en el último año.
- Un gráfico con los tuits escritos por semana durante el último año superpuesto a los mensajes escritos en Twitter en las mismas fechas hace 2 años.
- Un gráfico con los mensajes por semana escritos en Twitter durante el último año.
- Un gráfico con las variaciones porcentuales por semana y por intervalos de dos semanas de los mensajes escritos en Twitter y las consultas realizadas en Google.

Figura 32. Visualiación de la aplicación visual justo después de su creación



Después crearemos un informe en el que definiremos los datos que se mostrarán. Para ello, en el menú superior de Visual Studio seleccionamos la pestaña *Proyecto* y la subpestaña *Agregar nuevos elementos*. Visual Studio nos mostrará un cuadro de diálogo en el que deberemos seleccionar la opciones *Reporting* e *Informe*.

Figura 33. Creación de un informa para la aplicación



El siguiente paso consistirá en agregar las gráficas que debe contener nuestro cuadro de mando.

Para crear gráficas en Microsoft Visual Studio con los datos provenientes del almacén de datos, se deben recuperar los datos a partir de procedimientos almacenados que sigan la siguiente estructura:

```
create or replace function nombre_del_procedimiento return types.cursortype
as
   l_cursor types.cursorType;
begin
   open l_cursor for consulta_del_paso_anterior
end;
```

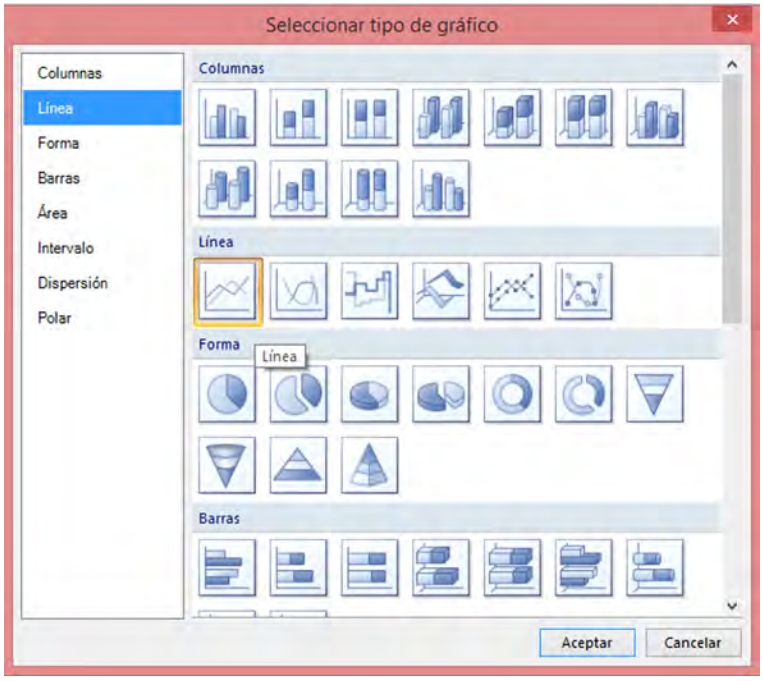
A modo de ejemplo, escribimos el código completo del procedimiento almacenado para recuperar el número de tuits semanales de los últimos 6 meses.

```
create or replace function proc_tweets_ultimos_6_meses return types.cursortype
as
   l_cursor  types.cursorType;
begin
   open l_cursor for
      SELECT SUM(t.num_mensajes) as tweets, f.semana as semana
      FROM H_TWEET t, DIM_FECHA f
      WHERE t.ID_FECHA = f.ID_FECHA and f.fecha > add_months( sysdate, -6)
      GROUP BY f.semana
      ORDER BY f.semana
end;
```

Una vez hayamos creado todos los procedimientos almacenados, podemos empezar con la generación de las gráficas.

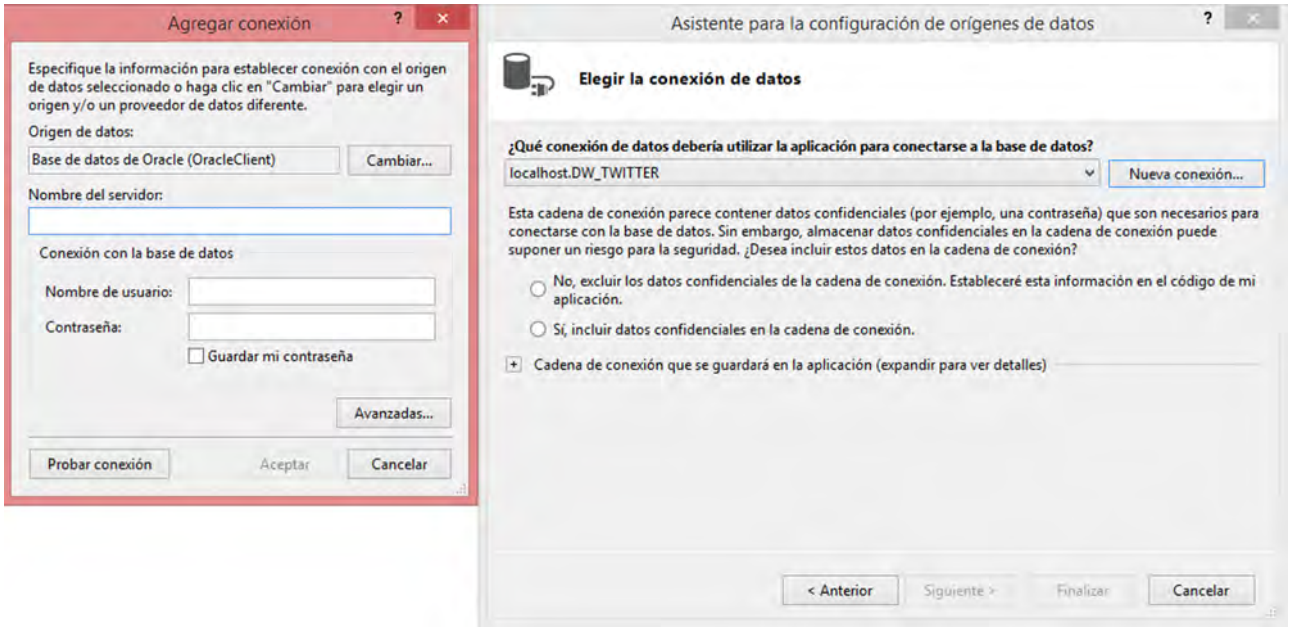
Para crear una nueva gráfica hacemos clic con el botón derecho del ratón en el área de trabajo del informe que estamos creando, escogemos las opciones *Insertar* y *Gráfica* y empezamos a crear el gráfico con el diálogo que mostrará Visual Studio.

Figura 34. Selección del tipo de gráfico a añadir en el informe



Si es la primera vez que creamos un gráfico en el proyecto nos pedirá crear la conexión a la *Base de datos*.

Figura 35. Selección de la base de datos

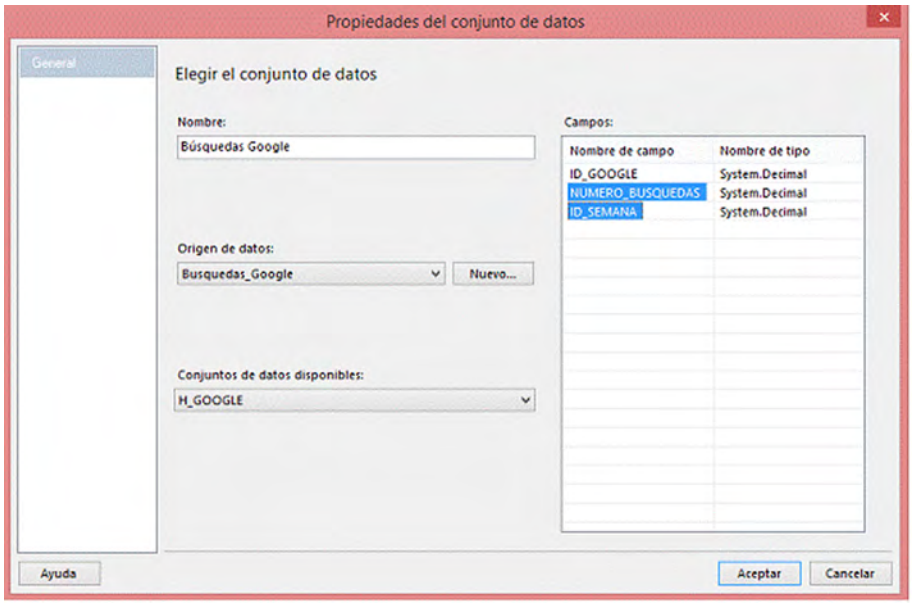


Una vez realizada la conexión, nos pedirá que seleccionemos las tablas de la conexión que deseamos tener disponibles en el proyecto. Las tablas que seleccionaremos para crear el reporte gráfico del cuadro de mandos serán todos los procedimientos almacenados que acabamos de crear y las tablas de los hechos y dimensiones. Se debe tener en cuenta que para crear el reporte gráfico solo

utilizaremos los procedimientos almacenados, pero para que estos funcionen correctamente también tendremos que seleccionar las tablas de datos que utilizan los procedimientos.

Con la conexión a la base datos creada y el conjunto de tablas y procedimientos almacenados que se utilizarán seleccionados, el diálogo nos pedirá seleccionar una tabla o procedimiento e indicar qué atributos queremos utilizar para generar la gráfica.

Figura 36. Selección de las tablas y atributos para los gráficos



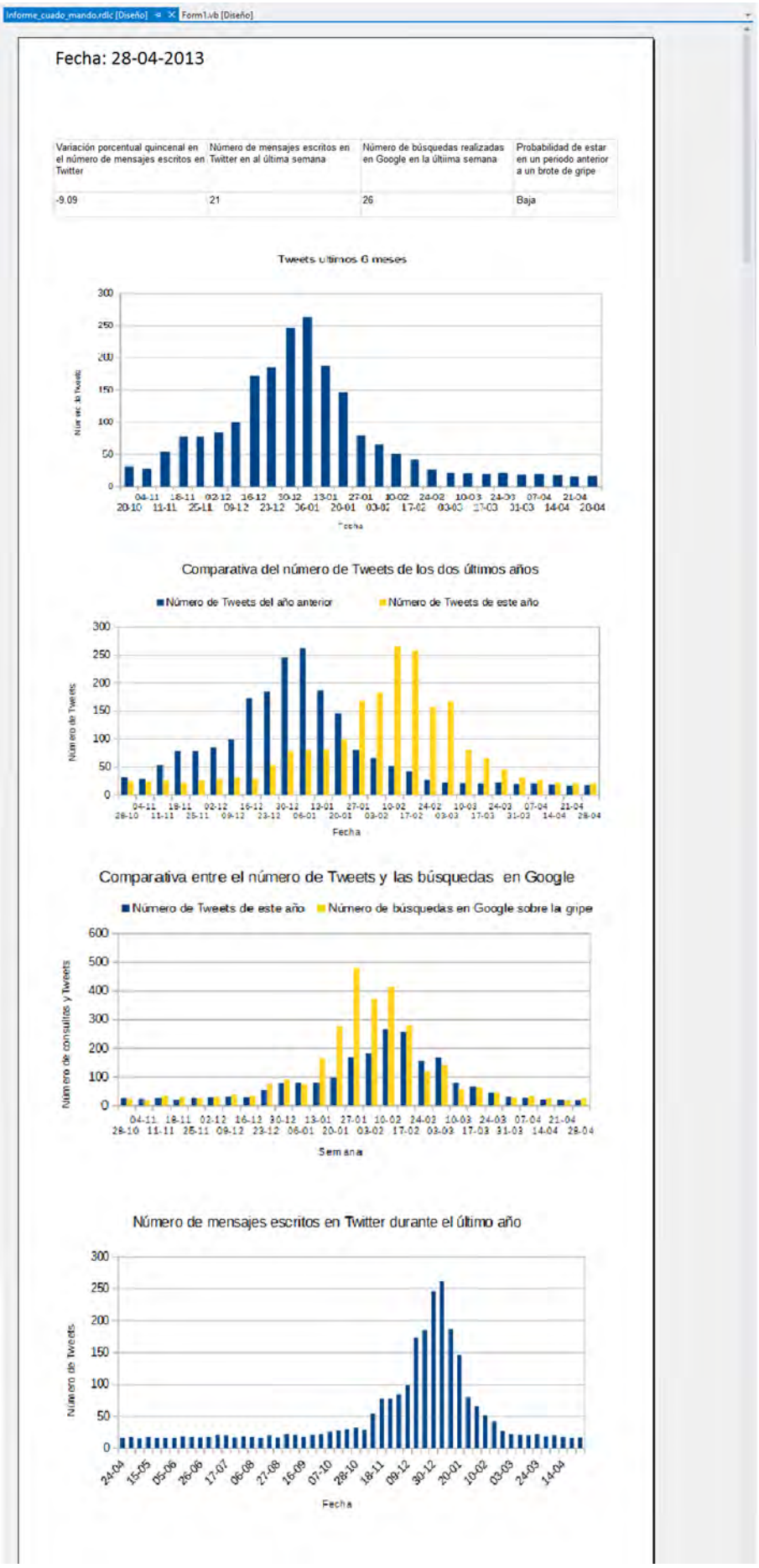
Después, para cada gráfico del cuadro de mando, ya podemos dar nombre a los ejes de abscisas y ordenadas e introducir la leyenda de datos.

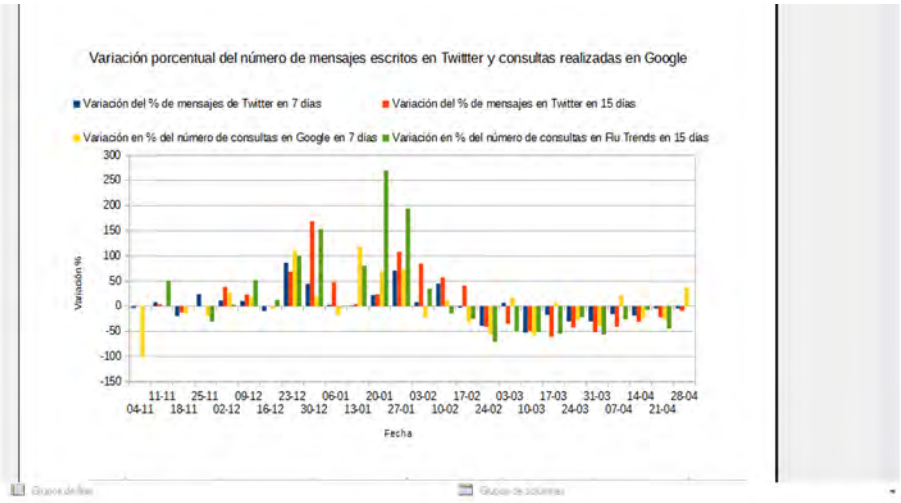
El último paso para finalizar la implementación del cuadro de mando es la generación de los indicadores que resumen la situación durante la última semana. Estos indicadores son:

- Variación porcentual quincenal en el número de mensajes escritos en Twitter.
- Número de mensajes escritos en Twitter en la última semana.
- Número de búsquedas realizadas en Google en la última semana.
- Probabilidad de estar en un periodo anterior a un brote de gripe.

El cuadro de mando resultante de la implementación se muestra en la figura siguiente. Podemos comprobar que el resultado muestra información sobre los indicadores definidos durante la fase de diseño del cuadro de mando.

Figura 37. Resultado de la implementación del cuadro de mandos




Variación porcentual del número de mensajes escritos en Twittter y consultas realizadas en Google
Variación del % de mensajes de Twitter en 7 días
Variación del % de mensajes en Twitter en 15 días
Variación en % del número de consultas en Google en 7 días
Variación en % del número de consultas en Flu Trends en 15 días
Variación %
300
250
200
150
100
50
0
-50
-100
-150
11-11 25-11 09-12 23-12 06-01 20-01 03-02 17-02 03-03 17-03 31-03 14-04 28-04
04-11 18-11 02-12 16-12 30-12 13-01 27-01 10-02 24-02 10-03 24-03 07-04 21-04
Fecha

Resumen

Este caso tenía por objetivo crear un sistema que dé soporte en la detección de nuevos brotes de gripe a partir Twitter.

Para ello se ha discutido cómo extraer la información de Twitter, se ha diseñado e implementado un cuadro de mando con un conjunto de indicadores que permitan detectar nuevos brotes de gripe, se ha verificado que los indicadores propuestos tienen capacidad predictiva y se ha creado un almacén de datos para poder tener actualizado el cuadro de mandos en todo momento con información actualizada.

En la primera parte del caso hemos definido el proceso para recuperar los datos de Twitter de forma automatizada. Posteriormente hemos definido los objetivos estratégicos que debía alcanzar el proyecto, hemos presentado un conjunto de indicadores que ayudan en la detección anticipada de un nuevo brote de gripe y hemos definido la apariencia del cuadro de mando.

A partir de los conjuntos de datos sobre las hospitalizaciones y sobre los mensajes escritos por los usuarios de Twitter en el año 2013, hemos testado que los indicadores propuestos tenían la capacidad de predecir la llegada de un nuevo brote de gripe con tres semanas de antelación. Para dar mayor robustez al sistema hemos incorporado datos sobre las búsquedas realizadas en Cataluña sobre Google relacionadas con la gripe. Aunque las fuentes de datos de Google y Twitter son completamente diferentes, permiten ver en qué magnitud los usuarios se interesan por la gripe en el transcurso del tiempo. Nuestra hipótesis inicial era que los datos estarían correlacionados y mostrarían una evolución similar en el transcurso del tiempo, como así ha sido.

Para implementar el proyecto, hemos diseñado e implementado un almacén de datos que permita almacenar los conjuntos de datos de interés. Se ha creado un conjunto de procesos ETL para poblar inicialmente el almacén de datos y para actualizarlo de forma periódica con los nuevos datos de Twitter y Google. Asimismo hemos definido un conjunto de consultas SQL que permiten recuperar, del almacén de datos, la información que se mostrará en el cuadro de mando.

El resultado final satisface el problema planteado, ya que ofrece información para estimar el advenimiento de un brote de gripe con antelación. Los resultados muestran que los datos de las redes sociales pueden integrarse en los sistemas de inteligencia de negocio de las empresas para tener más información del entorno donde operan las organizaciones. En el caso que nos ocupa, queda claro que el uso de las nuevas tecnologías de información nos permite prede-

cir de forma eficaz y con mayor antelación la llegada de un brote de gripe, lo que permite tener un margen de tiempo superior para organizar los recursos sanitarios e informar con mayor antelación a la población.