El servicio de texto completo

El servicio de búsqueda por texto completo tiene como objetivo mejorar la precisión y la velocidad de las consultas que se llevan a cabo sobre campos que contienen textos de grandes dimensiones: es decir, en las columnas de tipo char, nchar, varchar, nvarchar, varbinary y xml. Si la tabla tiene muchas líneas, una consulta con el operador like puede tardar varios minutos en ejecutarse. Si la columna tiene un índice por texto completo, la extracción de las líneas tarda sólo unos instantes.

La puesta en marcha de este servicio, utilizado para la indización, la consulta y la sincronización, necesita la presencia de una clave única (o clave primaria) en todas las tablas susceptibles de ser utilizadas en una búsqueda por texto completo. El índice por texto completo conserva una traza de todas las palabras significativas que se han empleado y de su ubicación. Para que la búsqueda sea acertada y rápida, sólo se deben indizar palabras relevantes y con sentido. Para identificar las palabras sin sentido, SQL Server utiliza una lista de palabras irrelevantes. Esta lista se conserva directamente en la base de datos. Sin embargo, como la incorporación de esta lista es específica de SQL Server 2008, en el caso de una migración desde SQL Server 2005, esta lista se conserva en forma de archivo externo.

Esta lista de palabras irrelevantes se puede modificar libremente para añadir palabras irrelevantes específicas de una empresa o de un contexto de trabajo. Por ejemplo, el nombre de la empresa puede considerarse como un palabra irrelevantes ya que es probable que aparezca con frecuencia.

El servicio de búsqueda por texto completo se basa en aspectos precisos para describir su puesta en marcha y su funcionamiento:

- Índice por texto completo: almacena la información relativa a las palabras significativas. Las búsquedas se efectúan a partir de esta información.
- Catálogo por texto completo: asociado a una instancia de SQL Server. Puede contener de 0 a n índices.
- Analizador léxico: en función del idioma y de sus reglas léxicas, el analizador léxico va a definir les fichas.
- Ficha: se trata de una cadena de caracteres (normalmente palabras) marcada por el analizador léxico.
- Generador de formas derivadas: en función del idioma, el generador de formas derivadas permite gestionar las diferentes formas que puede tomar un término, como por ejemplo, la conjugación de un verbo o bien la concordancia de un nombre o un adjetivo.
- Filtro: este elemento permite extraer el texto a partir de un archivo específico (.doc, por ejemplo) y registrar este texto en una columna de tipo varbinary(max), por ejemplo.
- Análisis o Alimentación: se trata del proceso que permite inicializar y de mantener actualizado el índice.
- Palabras irrelevantes: se trata de una lista de palabras que no tienen significado ni sentido para las búsquedas en modo texto. Esta lista de palabras es específica de cada idioma. El objetivo perseguido es reducir el número de palabras que tratar eliminando todas las palabras de unión, los artículos o los términos desprovistos de significado según el contexto. Por ejemplo el nombre de la empresa o un acrónimo utilizado habitualmente.

Con SQL Server 2008, el servicio de búsqueda por texto completo es soportado íntegramente por el servicio MSSQLServer.

Este servicio también está disponible para todas las bases contenidas en el servidor.

La activación de la utilización del servicio al nivel de la base ya no se hace con SQL Server 2008.

Implementación

La búsqueda lingüística en datos de tipo texto sólo es posible en las tablas activadas para la búsqueda por texto completo. La búsqueda lingüística, al contrario que el operador LIKE, que se basa en los caracteres, efectúa una comparación sobre las palabras y expresiones.

La puesta en marcha de una búsqueda por texto completo en una base de datos obliga a efectuar las operaciones siguientes:

- precisar las tablas y las columnas que deben inscribirse en la búsqueda por texto completo.
- realizar la indexación de los datos de las columnas inscritas y completar los índices por texto completo con las

palabras relevantes.

- ejecutar las consultas en las columnas inscritas para la búsqueda por texto completo.
- asegurarse de que todas las modificaciones realizadas en estas columnas se han propagado a los índices.

Estos índices no se completan en tiempo real, sino de manera asíncrona ya que:

- el tiempo de indexación es generalmente mucho más largo.
- las búsquedas por texto completo son, por regla general, bastante menos precisas que las búsquedas en modo estándar.

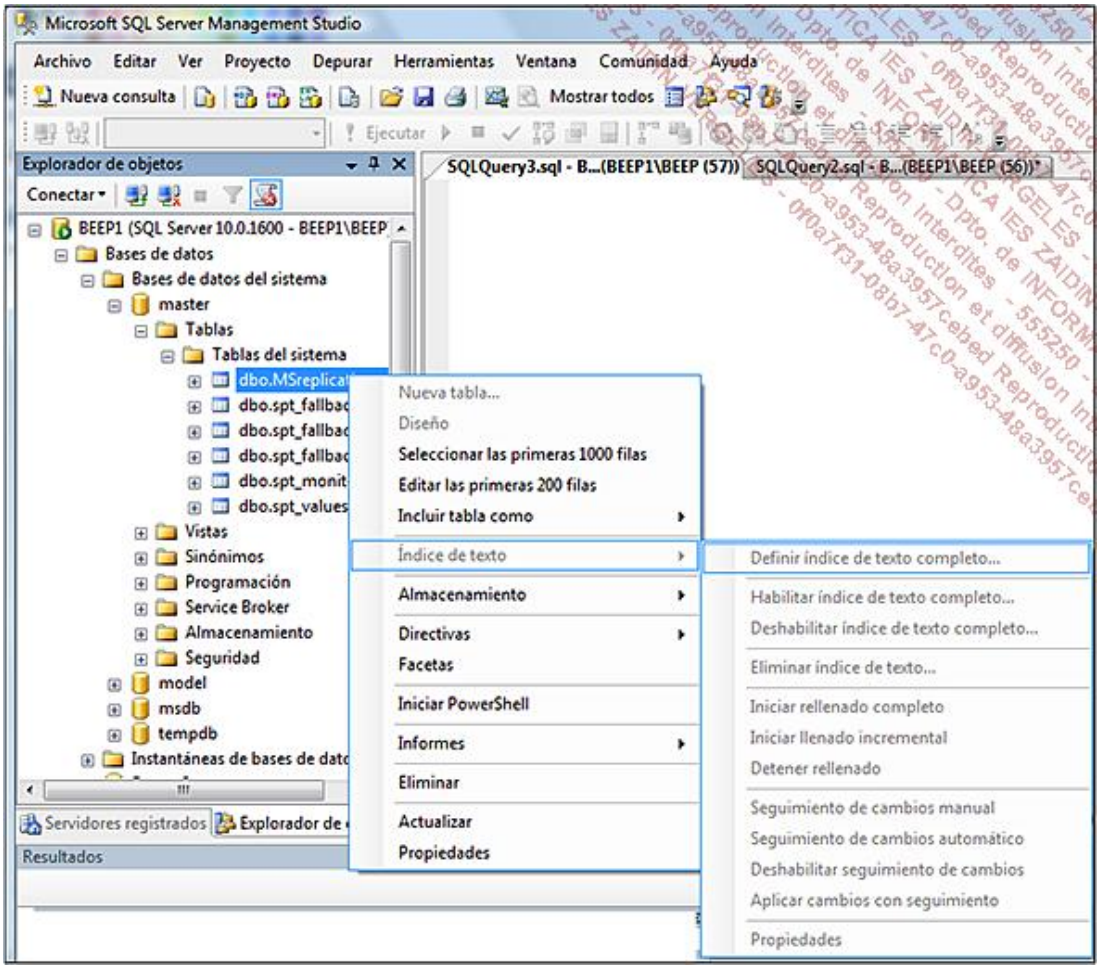
Puesta en marcha

La puesta en marcha del servicio de búsqueda por texto completo puede efectuarse desde SQL Server Management Studio en forma de scripts Transact SQL.

Las etapas del trabajo a realizar son:

- crear un catálogo;
- crear uno o varios índices que utilicen este catálogo;
- definir la lista de las palabras irrelevantes.

La puesta en marcha del servicio por texto completo puede realizarse tanto de manera gráfica desde SQL Server Management Studio, como por medio de scripts Transact SQL, o bien por medio del asistente de indexación por texto completo. Teniendo en cuenta que la creación de este tipo de índices es una operación puntual, la utilización del asistente puede ser muy útil. Se puede acceder desde el menú contextual asociado a la tabla en la que se va definir el índice.



Esta puesta en marcha también puede realizarse etapa por etapa por medio de comandos Transact SQL.

1. El catálogo

En el momento de su creación inicial, el índice por texto completo va a ser un consumidor importante en términos de lecturas y escrituras en el disco duro. Por lo tanto, es necesario que el índice sea definido en un sistema de archivos con alto rendimiento. El índice por texto completo se va definir por lo tanto en un catálogo que le sea propio. El catálogo se define obligatoriamente en la misma base que el índice. Así como un índice se define siempre en un catálogo, un catálogo puede contener uno o varios índices. Si la tabla indexada contiene muchas líneas, es recomendable que el índice por texto completo se defina en su propio catálogo. Por el contrario, si las tablas tienen un número razonable de líneas, entonces es posible definir un número razonable de índices en un mismo catálogo. Con el objetivo de que le catálogo esté optimizado, es recomendable agrupar en un mismo catálogo los índices que tienen una frecuencia de actualización parecida.

El catálogo será gestionado con las instrucciones Transact SQL CREATE FULLTEXT CATALOG, ALTER FULLTEXT CATALOG y DROP FULLTEXTCATALOG. Aquí se detalla únicamente la creación de un catálogo con la instrucción CREATE FULLTEXT CATALOG.

No es posible definir catálogos en las bases master, model y tempdb.

```
CREATE FULLTEXT CATALOG nombreCatálogo
  WITH ACCENT_SENSITIVITY = {ON|OFF}]
  [AS DEFAULT]
  [AUTHORIZATION nombrePropietario ]
```

`nombreCatálogo`

Nombre del catálogo. Cada catálogo tiene un nombre único. El nombre está limitado a 120 caracteres y respeta las reglas de nomenclatura de los identificadores en SQL Server. El nombre del catálogo será utilizado para nombrar al archivo. El nombre del archivo debe ser único.

`grupoDeArchivos`

Nombre del grupo de archivos al que va a pertenecer el catálogo.

`RutaRaíz`

Permite precisar el directorio que va a contener el archivo sysft_nombreCatálogo asociado al catálogo.

`ACCENT_SENSITIVITY`

Permite precisar si el catálogo debe distinguir o no los caracteres acentuados.

`AS DEFAULT`

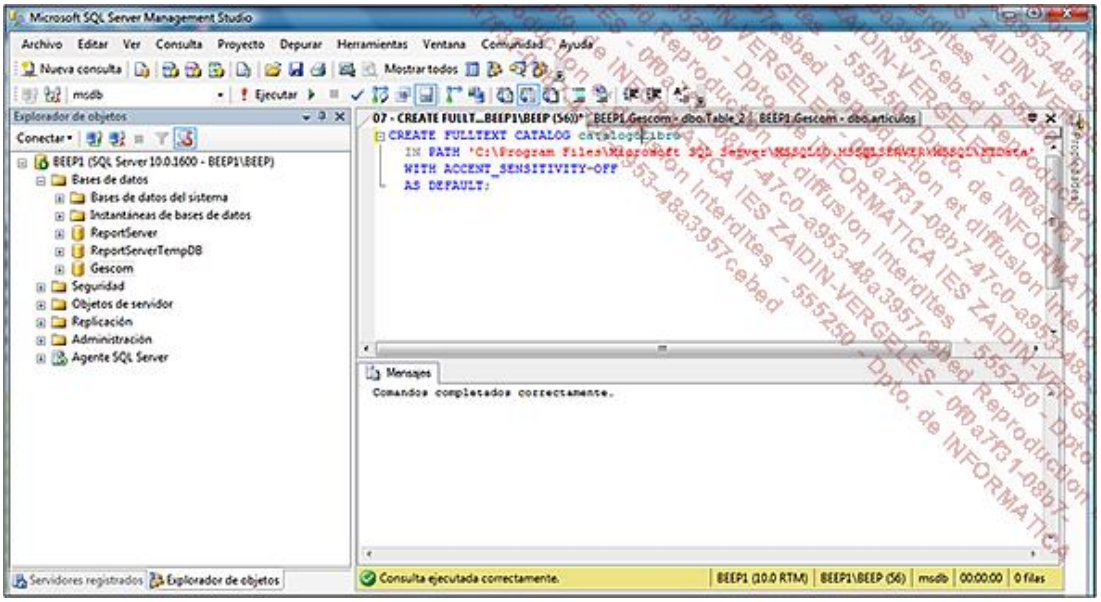
El catálogo se convierte en el catálogo por defecto para los índices por texto completo que se definan a partir de ese momento.

`nombrePropietario`

Si el creador del catálogo no es el propietario, es posible indicar el nombre del propietario.

Ejemplo

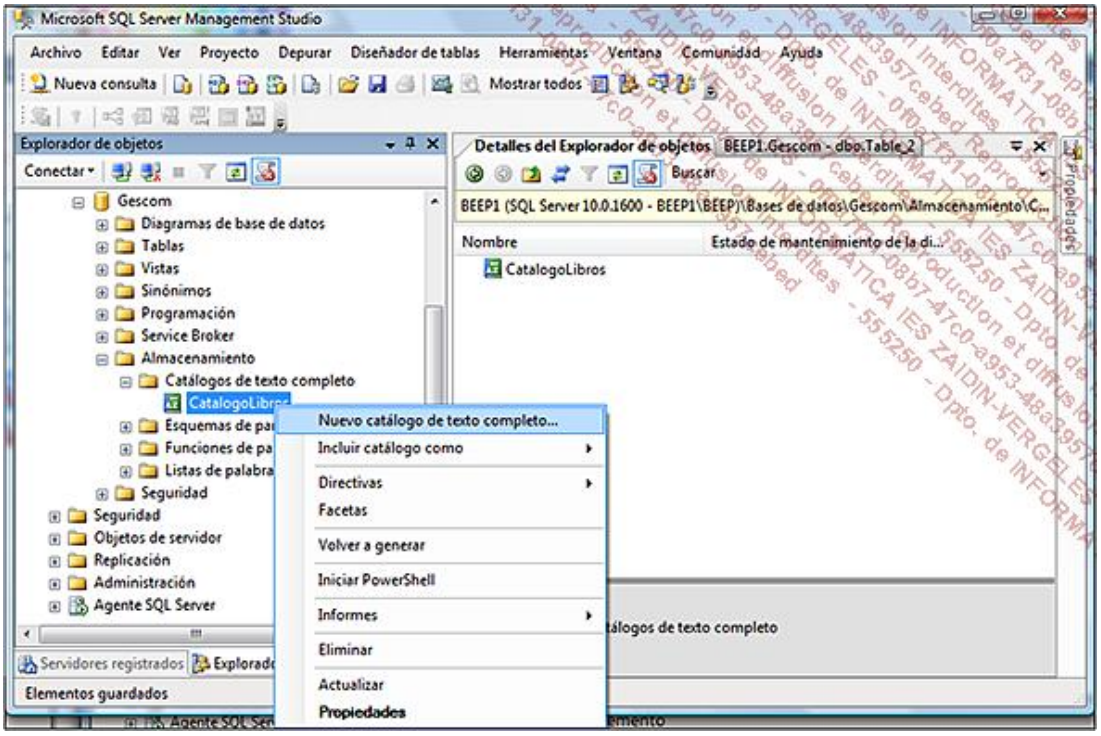
En el ejemplo siguiente, se define un catálogo de texto completo en la base Gescom con el directorio raíz C:\Program Files\Microsoft SQL Server\MSSQLIO.MSSQLCSERVER\MSSQL\FTData.



Las opciones ON FILEGROUP e IN PATH siguen presentes en la instrucción, pero no tienen ningún efecto cuando se trabaja en una instancia SQL Server 2008.

Para efectuar las operaciones de modificación y eliminación es necesario utilizar las instrucciones ALTER FULLTEXT CATALOG y DROP FULLTEXT CATALOG.

También es posible realizar estas operaciones desde la consola SQL Server Management Studio seleccionando la opción **Nuevo catálogo de texto completo** desde el menú contextual asociado a nodo **Almacenamiento - Catálogos de texto completo** en el explorador de objetos.



El índice por texto completo

El índice único utilizado por el índice por texto completo para hacer referencia a las líneas de información deberá ser lo más compacto posible con el objetivo de limitar la carga de trabajo. Un dato de tipo entero (int) encaja perfectamente.

El índice por texto completo se define con la instrucción CREATE FULLTEXT INDEX.

Este tipo de índice puede naturalmente definirse en columnas que contienen datos de tipo texto, pero también en las columnas de tipo varbinary(max) que almacenan textos directamente en un formato de datos específico (por ejemplo .doc). Este tipo de índice también se puede crear en las columnas de tipo xml.

```
CREATE FULLTEXT INDEX ON nombreTabla
     [(nombreColumna [TYPE COLUMN tipoDocumento]
          [LANGUAGE indicadorDeIdioma] [,...])]
     KEY INDEX nombreÍndice
          [ON nombreCatálogo]
     [WITH CHANGE_TRACKING
       {MANUAL | AUTO | OFF [, NO POPULATION]}]
[,STOPLIST={OFF|SYSTEM|nombreListaPalabrasIrrelevantes}]
```

`nombreTabla`

Se trata del nombre de la tabla en la que se ha definido el índice por texto completo.

`nombreColumna`

Nombre de la columna o de las columnas que participan en el índice.

`tipoDocumento`

Cuando la columna indexada contiene un documento, esta columna permite conocer el tipo de documento.

`indicadorDeIdioma`

Permite especificar el idioma en el que se guarda la información en la columna indexada. Este indicador sólo es útil en caso de que el idioma utilizado para almacenar la información difiera del idioma por defecto definido en SQL Server. Por ejemplo, para indexar una columna que contiene textos en inglés cuando SQL Server está configurado con español como idioma por defecto.

`nombreÍndice`

Nombre del índice único que será utilizado para hacer referencia a las líneas de información en la tabla.

`nombreCatálogo`

Nombre del catálogo utilizado por el índice. Si no se especifica ningún nombre de catálogo, se utiliza el catálogo por defecto (el creado con la cláusula AS DEFAULT).

`WITH CHANGE_TRACKING`

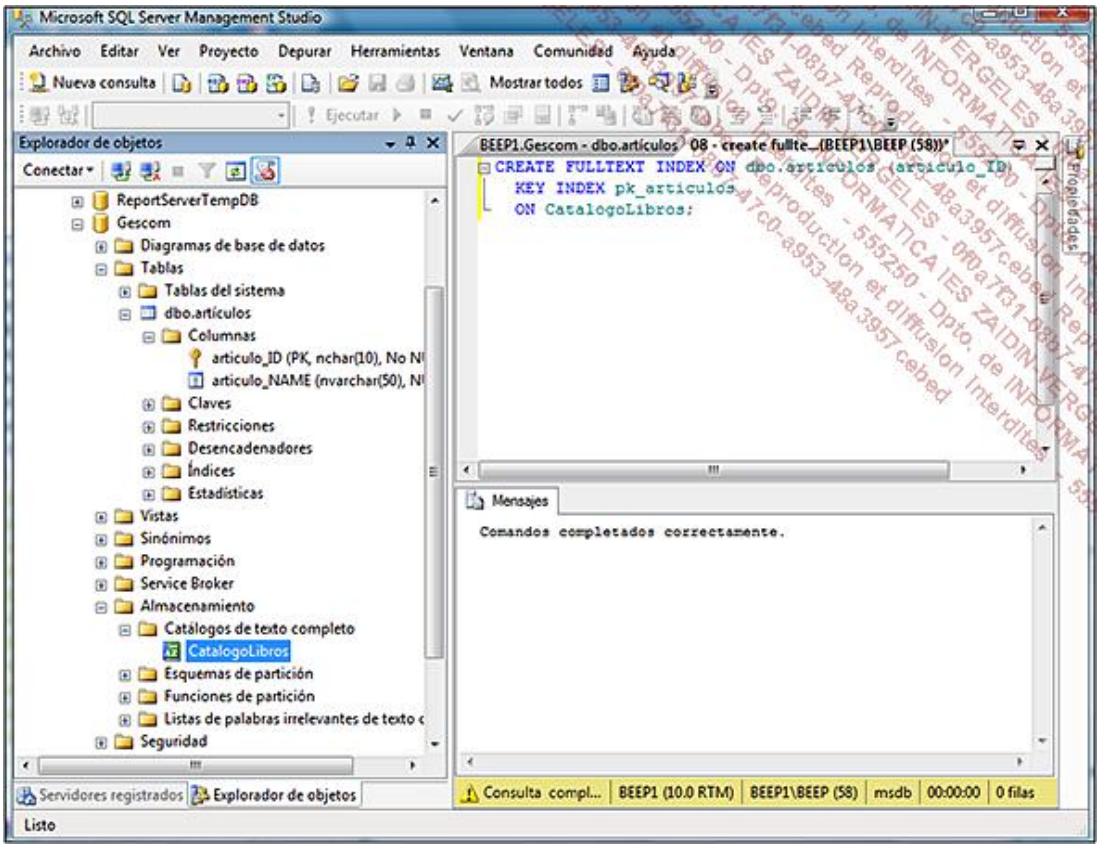
Esta cláusula permite especificar cómo se informa al índice de las modificaciones realizadas en las columnas indexadas.

`STOPLIST`

Esta opción permite precisar la lista de palabras irrelevantes asociadas al índice: si no hay ninguna (OFF), si es la lista por defecto (SYSTEM), o si es una lista personalizada, en cuyo caso es necesario indicar el nombre.

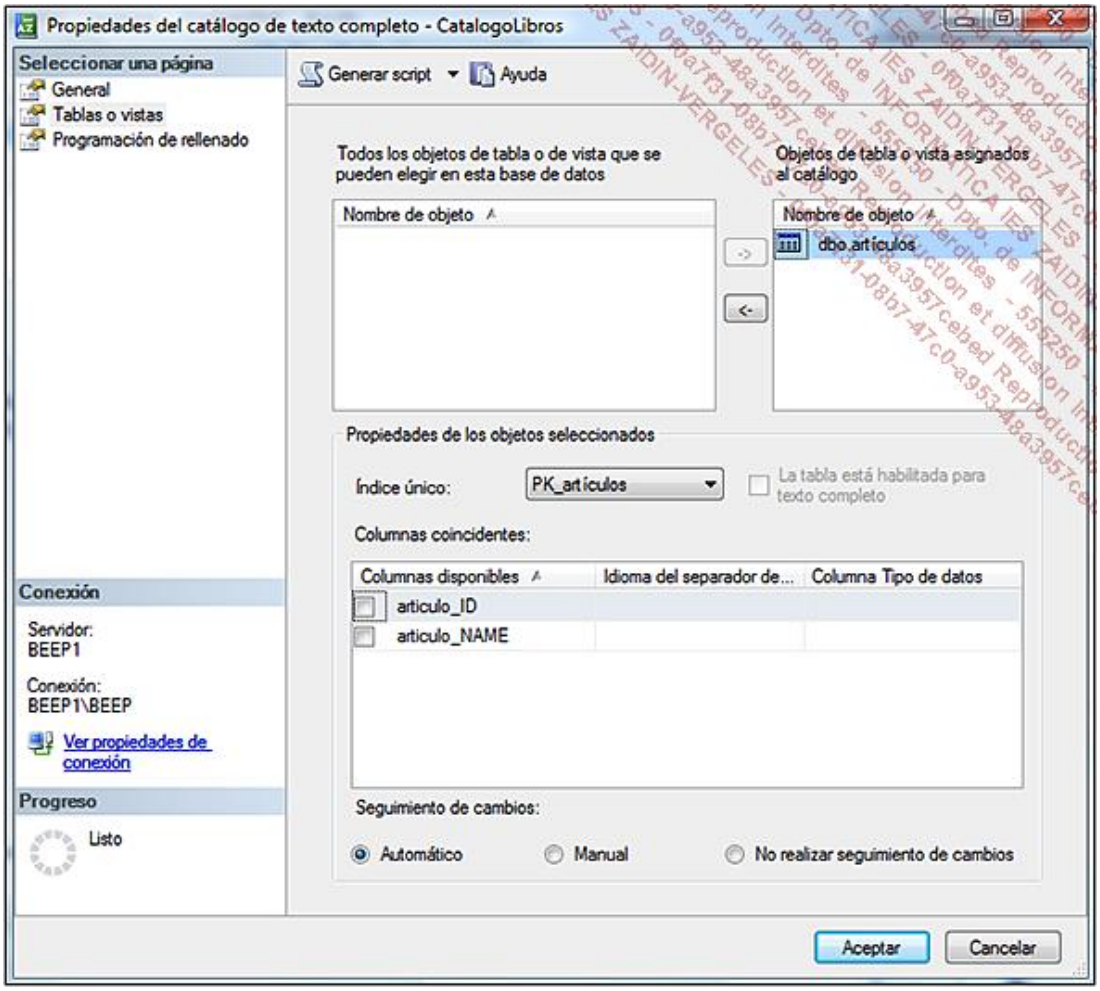
Ejemplo

En el ejemplo a continuación, la columna articulo_ID de la tabla de los artículos, está indexada. Este índice utiliza el catálogo ***CatalogoLibros*** *y realiza el seguimiento de las modificaciones de manera automática.*



Desde SQL Server Management Studio, es posible visualizar los detalles de este índice por medio de las propiedades del catálogo.

Las propiedades del catálogo están accesibles al seleccionar **Propiedades** desde el menú contextual asociado al catálogo.



2. La lista de palabras irrelevantes

Esta lista permite definir cuáles son las palabras vacías de significado y que por lo tanto no deben tenerse en cuenta en el índice.

Esta lista de palabras irrelevantes está disponible únicamente desde SQL Server 2008. Por lo tanto, la base de datos debe estar a un nivel 100 de compatibilidad para poder utilizarla.

Sintaxis

```
CREATE FULLTEXT STOPLIST nombreListaPalabrasIrrelevantes
[FROM {nombreListaPalabrasIrrelevantesFuente|SYSTEM STOPLIST}]
[AUTHORIZATION propietario];
```

`nombreListaPalabrasIrrelevantes`

Se trata del nombre asignado a la lista de palabras irrelevantes que está siendo creada.

`nombreListaPalabrasIrrelevantesFuente`

Se trata del identificador de una lista de palabras irrelevantes cuyo contenido se va a copiar en la lista de palabras que está siendo creada.

`SYSTEM STOPLIST`

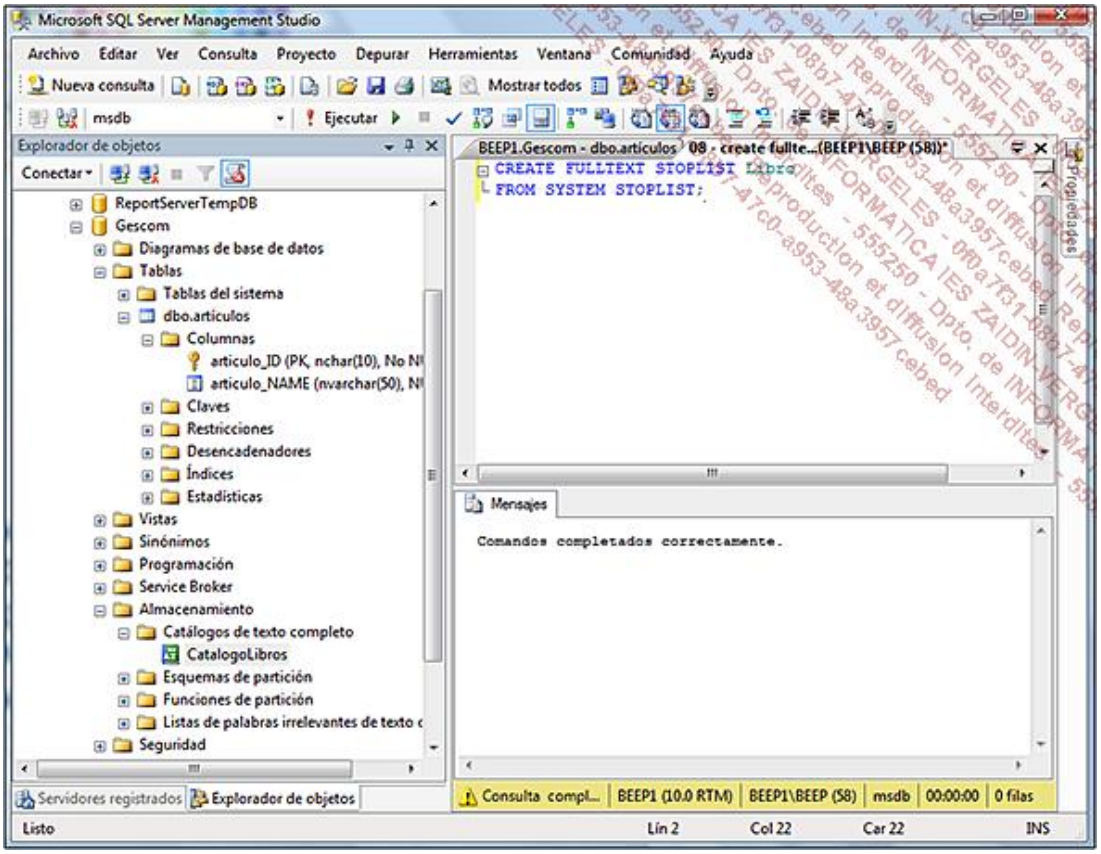
Con esta opción, la lista de las palabras irrelevantes se define a partir de la lista situada en la base de datos de recursos.

`propietario`

Se trata esta vez de especificar explícitamente el propietario de esta lista. Esta opción sólo es necesaria en caso de que la persona que ejecute el comando no sea el propietario del índice.

Ejemplo:

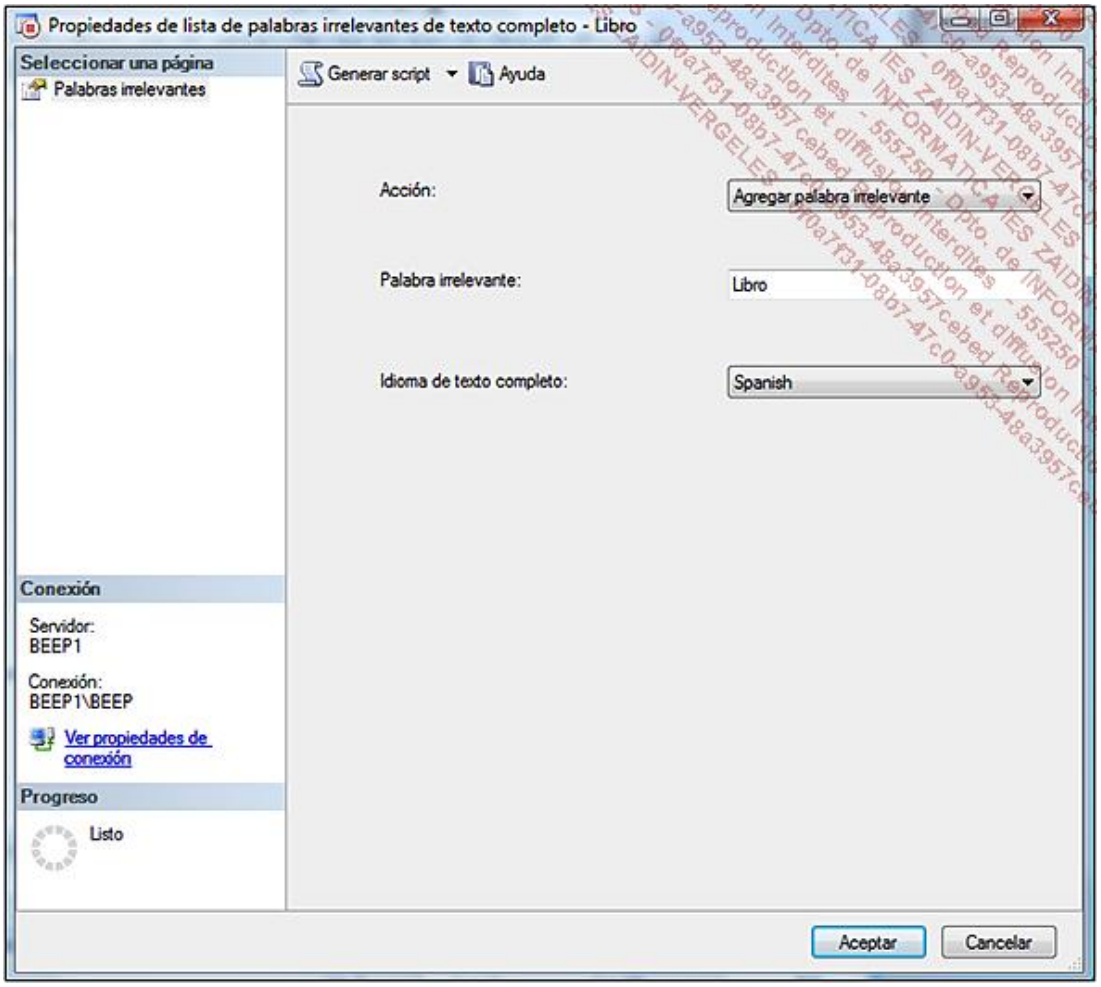
El ejemplo siguiente muestra la creación de una lista de palabras irrelevantes desde un script Transact SQL.



Esta lista también puede gestionarse desde SQL Server Management Studio a partir del nodo **Almacenamiento - Lista de palabras irrelevantes de texto completo**.

Ejemplo

Se añade a la lista la palabra Libro.



Una vez definida, esta lista de palabras irrelevantes puede modificarse con la ayuda de la instrucción ALTER FULLTEXT STOPLIST y eliminarse con DROP FULLTEXT STOPLIST. El comando ALTER permite modificar la lista tanto para añadir como para eliminar palabras desprovistas de significado.

Sintaxis

```
ALTER FULLTEXT STOPLIST listaPalabrasIrrelevantes
ADD 'nuevaPalabra'  LANGUAGE idioma;
```

`listaPalabrasIrrelevantes`

Representa el identificador de la lista que se va a modificar.

`nuevaPalabra`

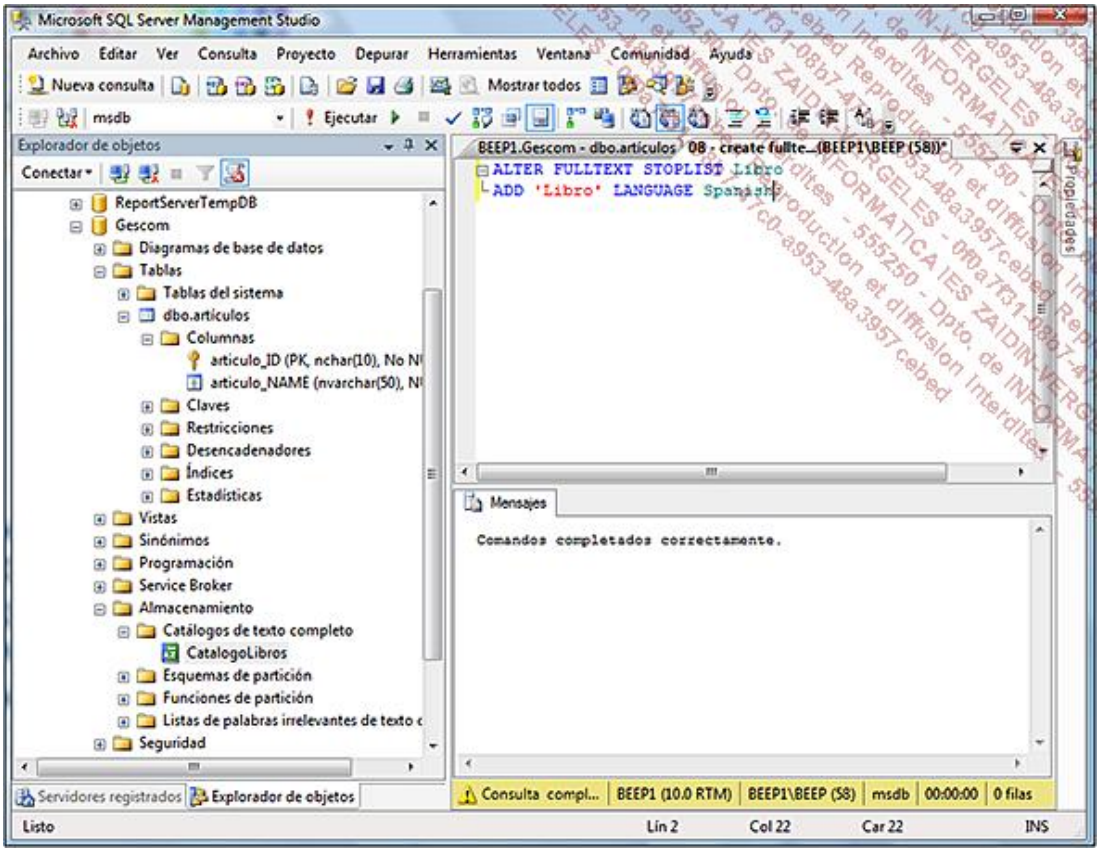
Corresponde al nuevo término vacío de significado. Los términos añadidos representan palabras específicas de un vocabulario determinado que están presentes muy habitualmente para hacer eficaz la indexación.

`idioma`

Representa el idioma para el que se define esta palabra. La codificación de los diferentes idiomas registrados en SQL Server está disponible consultando la tabla **sys.syslanguages**.

Ejemplo

En el ejemplo siguiente, la palabra Libro se considera desprovista de significado para el idioma español.



Inicializar el índice

Después de crear el índice, es necesario enriquecerlo con datos. Esta operación no se realiza de manera instantánea, ya que puede requerir una fuerte carga de trabajo en el lado del servidor. Por lo tanto, es preferible planificar el rellenado del índice para una época de poca actividad.

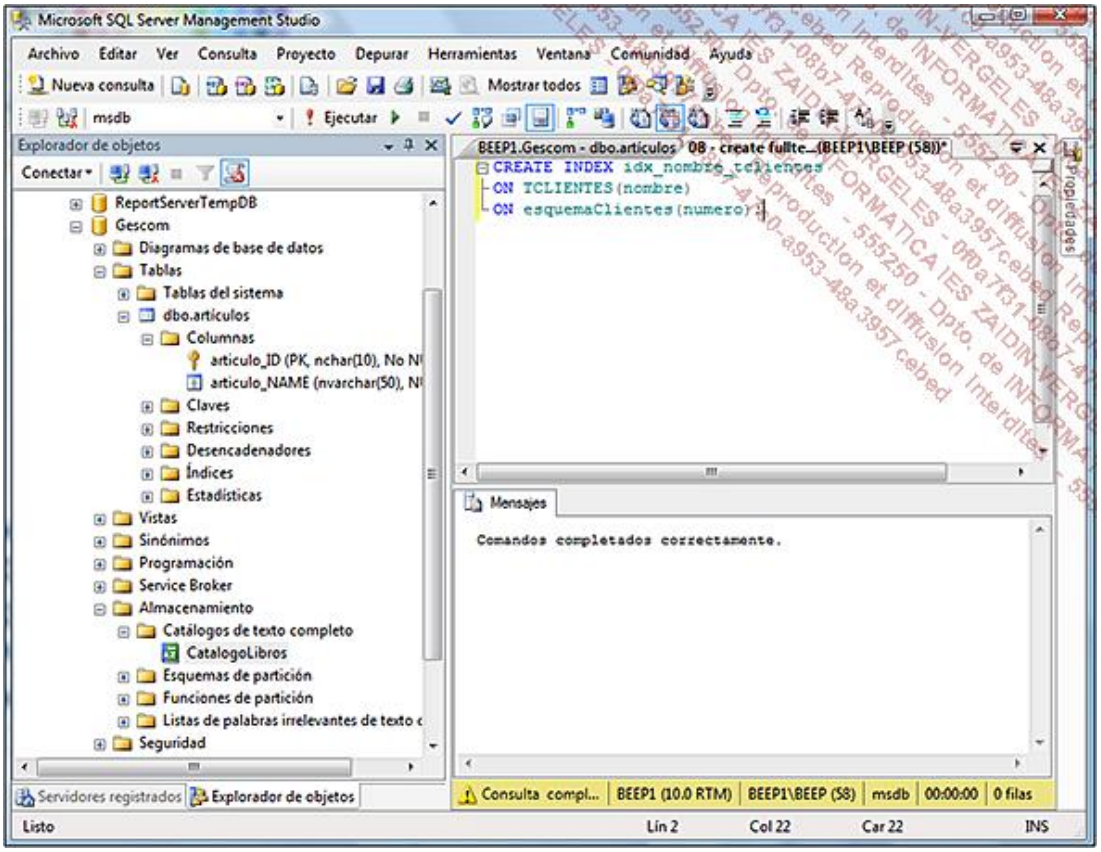
Existen tres maneras diferentes de inicializar y mantener actualizado el índice completo: en función de las modificaciones o de forma periódica.

La inicialización completa del índice es, sin duda, la manera la más natural de trabajar con los índices, ya que se indexan todos los términos de manera global, ya sea en el momento de la creación del índice, o bien basándose en un incremento temporal determinado.

Con el modo de rellenado basado en las modificaciones, SQL Server guarda una traza de todos los datos añadidos o modificados. El informe de estas modificaciones hacia el índice de texto completo puede efectuarse de manera manual, utilizando un script con una ejecución planificada, o bien de manera continua.

Por último, el rellenado basado en incrementos temporales se basa en un valor de tipo timestamp. Cuando se añade información a la tabla, la columna de tipo timestamp toma valor. Si la tabla no tiene una columna de tipo timestamp, no es posible utilizar este tipo de rellenado. Al final del proceso de rellenado, el valor timestamp actual se conserva en los metadatos, de manera que en el próximo proceso de rellenado, será posible tener en cuenta únicamente los datos más recientes.

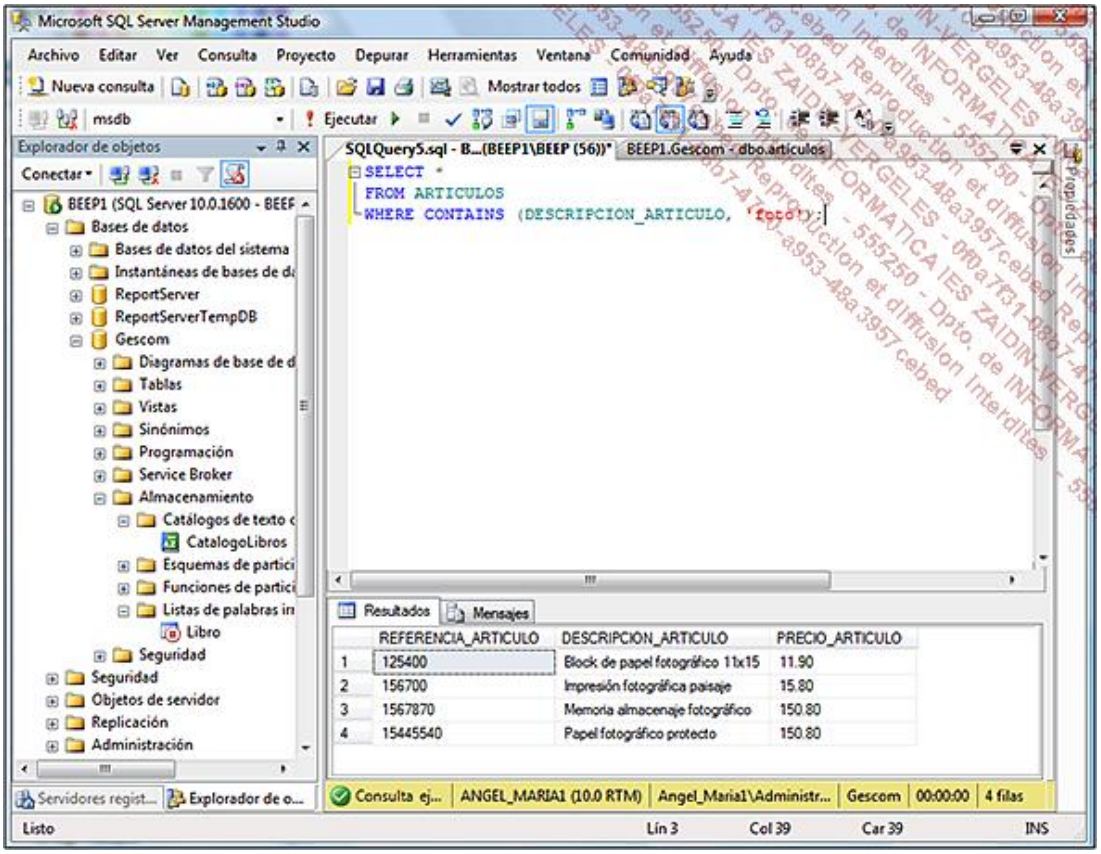
Es posible definir esta operación a partir de la ventana de propiedades del catálogo, como se ilustra en la pantalla siguiente.



Esta operación también puede realizarse en Transact SQL con la instrucción `ALTER FULLTEXT INDEX ON` *`nombreTabla START FULL POPULATION`* para una inicialización completa. Esta instrucción tiene diferentes opciones que permiten, por ejemplo, activar o desactivar el índice.

Utilización dentro de las consultas

La utilización de los índices de texto completo se efectúa por medio de dos predicados, **CONTAINS** y **FREETEXT**, que pueden utilizarse en cualquier cláusula WHERE. También existen dos funciones Transact SQL que combinan un conjunto de líneas y pueden utilizarse en una cláusula FROM de una consulta SELECT, a saber, **CONTAINSTABLE** y **FREETEXTTABLE**.



3. Encontrar la información relativa a los índices de texto completo

Toda la información relativa a estos índices puede encontrarse consultando las diferentes vistas del catálogo de sistema.

Las tres vistas que se presentan a continuación son las más frecuentemente utilizadas.

sys.fulltext_index_catalog_usages

Permite obtener la lista de los catálogos definidos.

sys.fulltext_index_column

Permite identificar las columnas que participan en un índice de texto completo.

sys_dm_fts_index_population

Permite obtener la información de rellenado en los índices activos de tipo texto completo.