عنوان: ایجاد کاتالوگ‌های Full text search و ایندکس‌های آن
نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۱۲/۱۶ ۰:۲۶
آدرس: www.dotnettips.info
گروه‌ها: SQL Server

جستجو بر روی خواص و متادیتای اسناد آفیس

همانطور که در قسمت قبل نیز عنوان شد، فیلترهای FTS آفیس، علاوه بر اینکه امکان جستجوی پیشرفته FTS را بر روی کلیه فایل‌های مجموعه آفیس میسر می‌کنند، امکان جستجوی FTS را بر روی خواص ویژه اضافی آن‌ها، مانند نام نویسنده، واژه‌های کلیدی، تاریخ ایجاد و امثال آن نیز به همراه دارند.
اینکه چه خاصیتی را بتوان جستجو کرد نیز بستگی به نوع فیلتر نصب شده دارد. برای تعریف خواص قابل جستجوی یک سند، باید یک SEARCH PROPERTY LIST را ایجاد کرد:

```
CREATE SEARCH PROPERTY LIST WordSearchPropertyList;
GO

ALTER SEARCH PROPERTY LIST WordSearchPropertyList
   ADD 'Authors'
   WITH (PROPERTY_SET_GUID = 'F29F85E0-4FF9-1068-AB91-08002B27B3D9',
             PROPERTY_INT_ID = 4,
             PROPERTY_DESCRIPTION = 'System.Authors - authors of a given item.');
GO
```

در این تعریف، PROPERTY_INT_ID و PROPERTY_SET_GUIDها استاندارد بوده و لیست آن‌ها را در آدرس ذیل می‌توانید مشاهده نمائید:

Find Property Set GUIDs and Property Integer IDs for Search Properties

بهبود کیفیت جستجو توسط Stop lists و Stop words

به یک سری از کلمات و حروف، اصطلاحا noise words گفته می‌شود. برای مثال در زبان انگلیسی حروف و کلماتی مانند a، is، the و and به صورت خودکار از FTS حذف می‌شوند؛ چون جستجوی آن‌ها بی‌حاصل است. به این‌ها stop words نیز می‌گویند.
با استفاده از کوئری ذیل می‌توان لیست stop words تعریف شده در بانک اطلاعاتی جاری را مشاهده کرد:

```
-- Check the Stopwords list
SELECT w.stoplist_id,
   l.name,
   w.stopword,
   w.language
FROM sys.fulltext_stopwords AS w
   INNER JOIN sys.fulltext_stoplists AS l
     ON w.stoplist_id = l.stoplist_id;
```

و برای تعریف stop words از دستورات ذیل کمک گرفته می‌شود:

```
-- Stopwords list
CREATE FULLTEXT STOPLIST SQLStopList;
GO

-- Add a stopword
ALTER FULLTEXT STOPLIST SQLStopList
 ADD 'SQL' LANGUAGE 'English';
GO
```

کاتالوگ‌های Full Text Search

ایندکس‌های ویژه‌ی FTS، در مکان‌هایی به نام Full Text Catalogs ذخیره می‌شوند. این کاتالوگ‌ها صرفا یک شیء مجازی بوده و تنها برای تعریف ظرفی دربرگیرنده‌ی ایندکس‌های FTS تعریف می‌شوند. در نگارش‌های پیش از 2012 اس کیوال سرور، این کاتالوگ‌ها اشیایی فیزیکی بودند؛ اما اکنون تبدیل به اشیایی مجازی شده‌اند.
حالت کلی تعریف یک fulltext catalog به نحو ذیل است:

```
create fulltext catalog catalog_name
on filegroup filegroup_name
in path  'rootpath'
with some_options
as default
authoriztion owner_name
accent_sensivity = {on|off}
```

اما اکثر گزینه‌های آن مانند on filegroup و in path صرفا برای حفظ سازگاری با نگارش‌های قبلی حضور دارند و دیگر نیازی به ذکر آن‌ها نیست؛ چون تعریف کننده‌ی ماهیت فیزیکی این کاتالوگ‌ها می‌باشند.
به صورت پیش فرض حساسیت به لهجه یا accent_sensivity خاموش است. اگر روشن شود، باید کل ایندکس مجددا بازسازی شود.

ایجاد ایندکس‌های Full Text

پس از ایجاد یک fulltext catalog، اکنون نوبت به تعریف ایندکس‌هایی فیزیکی هستند که داخل این کاتالوگ‌ها ذخیره خواهند شد:

```
-- Full-text catalog
CREATE FULLTEXT CATALOG DocumentsFtCatalog;
GO

-- Full-text index
CREATE FULLTEXT INDEX ON dbo.Documents
(
  docexcerpt Language 1033,
  doccontent TYPE COLUMN doctype
  Language 1033
  STATISTICAL_SEMANTICS
)
KEY INDEX PK_Documents
ON DocumentsFtCatalog
WITH STOPLIST = SQLStopList,
  SEARCH PROPERTY LIST = WordSearchPropertyList,
  CHANGE_TRACKING AUTO;
GO
```

در اینجا توسط KEY INDEX نام منحصربفرد ایندکس مشخص می‌شود.
CHANGE_TRACKING AUTO به این معنا است که SQL Server به صورت خودکار کار به روز رسانی این ایندکس را با تغییرات رکوردها انجام خواهد داد.
ذکر STATISTICAL_SEMANTICS، منحصر به SQL Server 2012 بوده و کار آن تشخیص واژه‌های کلیدی و ایجاد ایندکس‌های یافتن اسناد مشابه است. برای استفاده از آن حتما نیاز است مطابق توضیحات قسمت قبل، Semantic Language Database پیشتر نصب شده باشد.
توسط STOPLIST، لیست واژه‌هایی که قرار نیست ایندکس شوند را معرفی خواهیم کرد. SQLStopList را در ابتدای بحث ایجاد کردیم.
1033 Language به معنای استفاده از زبان US English است.
نحوه‌ی استفاده از SEARCH PROPERTY LIST ایی که پیشتر تعریف کردیم را نیز در اینجا ملاحظه می‌کنید.

مثالی برای ایجاد ایندکس‌های FTS

برای اینکه ربط منطقی نکات عنوان شده را بهتر بتوانید بررسی و آزمایش کنید، مثال ذیل را درنظر بگیرید.

ابتدا جدول Documents را برای ذخیره سازی تعدادی سند، ایجاد می‌کنیم:

```
CREATE TABLE dbo.Documents
(
  id INT IDENTITY(1,1) NOT NULL,
  title NVARCHAR(100) NOT NULL,
  doctype NCHAR(4) NOT NULL,
  docexcerpt NVARCHAR(1000) NOT NULL,
  doccontent VARBINARY(MAX) NOT NULL,
  CONSTRAINT PK_Documents
PRIMARY KEY CLUSTERED(id)
);
```

اگر به این جدول دقت کنید، هدف از آن ذخیره‌ی اسناد آفیس است که فیلترهای FTS آن‌را در قسمت قبل نصب کردیم. ستون doctype، معرف نوع سند و doccontent ذخیره کننده‌ی محتوای کامل سند خواهند بود.

سپس اطلاعاتی را در این جدول ثبت می‌کنیم:

```
-- Insert data
-- First row
INSERT INTO dbo.Documents
(title, doctype, docexcerpt, doccontent)
SELECT N'Columnstore Indices and Batch Processing',
 N'docx',
 N'You should use a columnstore index on your fact tables,
   putting all columns of a fact table in a columnstore index.
   In addition to fact tables, very large dimensions could benefit
   from columnstore indices as well.
   Do not use columnstore indices for small dimensions. ',
 bulkcolumn
FROM OPENROWSET
 (BULK 'C:\Users\Vahid\Desktop\Updates\fts_docs\ColumnstoreIndicesAndBatchProcessing.docx',
  SINGLE_BLOB) AS doc;

-- Second row
INSERT INTO dbo.Documents
(title, doctype, docexcerpt, doccontent)
SELECT N'Introduction to Data Mining',
 N'docx',
 N'Using Data Mining is becoming more a necessity for every company
   and not an advantage of some rare companies anymore. ',
 bulkcolumn
FROM OPENROWSET
 (BULK 'C:\Users\Vahid\Desktop\Updates\fts_docs\IntroductionToDataMining.docx',
  SINGLE_BLOB) AS doc;

-- Third row
INSERT INTO dbo.Documents
(title, doctype, docexcerpt, doccontent)
SELECT N'Why Is Bleeding Edge a Different Conference',
 N'docx',
 N'During high level presentations attendees encounter many questions.
   For the third year, we are continuing with the breakfast Q&A session.
   It is very popular, and for two years now,
   we could not accommodate enough time for all questions and discussions! ',
 bulkcolumn
FROM OPENROWSET
 (BULK 'C:\Users\Vahid\Desktop\Updates\fts_docs\WhyIsBleedingEdgeADifferentConference.docx',
  SINGLE_BLOB) AS doc;

-- Fourth row
INSERT INTO dbo.Documents
(title, doctype, docexcerpt, doccontent)
SELECT N'Additivity of Measures',
 N'docx',
 N'Additivity of measures is not exactly a data warehouse design problem.
   However, you have to realize which aggregate functions you will use
   in reports for which measure, and which aggregate functions
   you will use when aggregating over which dimension.',
 bulkcolumn
FROM OPENROWSET
 (BULK 'C:\Users\Vahid\Desktop\Updates\fts_docs\AdditivityOfMeasures.docx',
  SINGLE_BLOB) AS doc;
GO
```

4 ردیف ثبت شده در جدول اسناد، نیاز به 4 فایل docx نیز دارند که آن‌ها را از آدرس ذیل می‌توانید برای تکمیل ساده‌تر آزمایش دریافت کنید:

fts_docs.zip

در ادامه می‌خواهیم قادر باشیم تا بر روی متادیتای نویسنده‌ی این اسناد نیز جستجوی کامل FTS را انجام دهیم. به همین جهت SEARCH PROPERTY LIST آن‌را نیز ایجاد خواهیم کرد:

```
-- Search property list
CREATE SEARCH PROPERTY LIST WordSearchPropertyList;
GO
ALTER SEARCH PROPERTY LIST WordSearchPropertyList
 ADD 'Authors'
 WITH (PROPERTY_SET_GUID = 'F29F85E0-4FF9-1068-AB91-08002B27B3D9',
PROPERTY_INT_ID = 4,
PROPERTY_DESCRIPTION = 'System.Authors - authors of a given item.');
GO
```

همچنین می‌خواهیم از واژه‌ی SQL در این اسناد، در حین ساخت ایندکس‌های FTS صرفنظر شود. برای این منظور یک FULLTEXT STOPLIST را به نام SQLStopList ایجاد کرده و سپس واژه‌ی مدنظر را به آن اضافه می‌کنیم:

```
-- Stopwords list
CREATE FULLTEXT STOPLIST SQLStopList;
GO
-- Add a stopword
ALTER FULLTEXT STOPLIST SQLStopList
 ADD 'SQL' LANGUAGE 'English';
GO
```

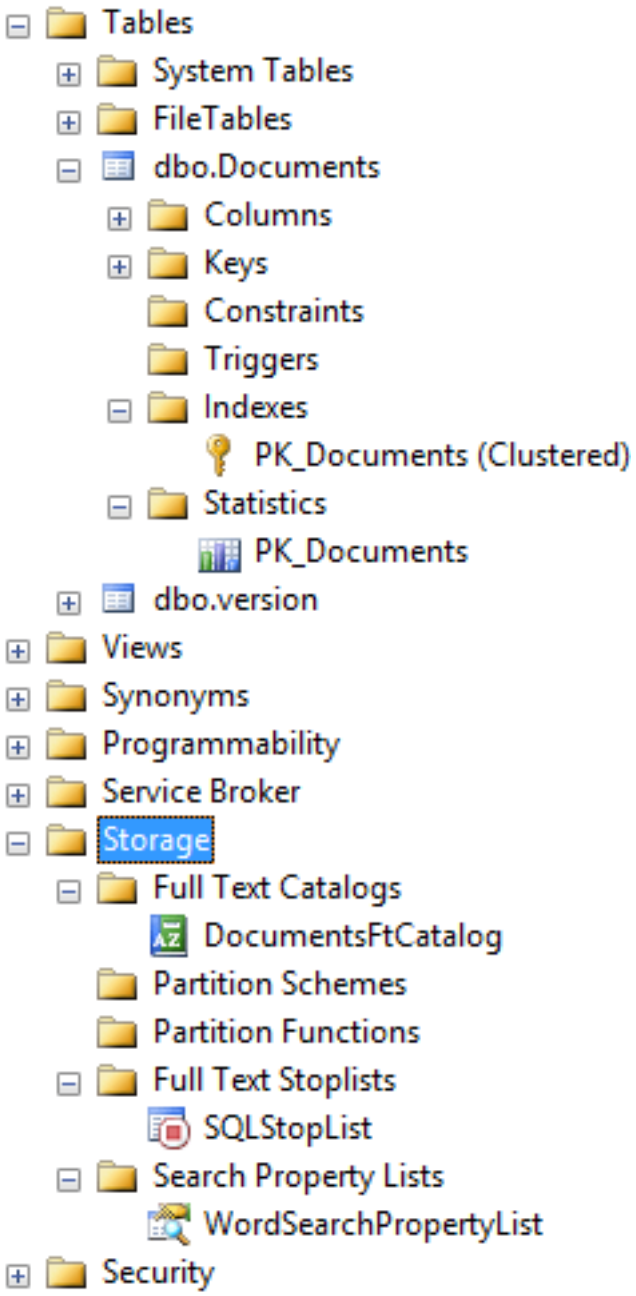
صحت عملیات آن‌را توسط کوئری «Check the Stopwords list» ذکر شده در ابتدای بحث می‌توانید بررسی کنید.

اکنون زمان ایجاد یک کاتالوگ FTS است:

```
-- Full-text catalog
CREATE FULLTEXT CATALOG DocumentsFtCatalog;
GO
```

با توجه به اینکه در نگارش‌های جدید SQL Server این کاتالوگ صرفا ماهیتی مجازی دارد، ساده‌ترین Syntax آن برای کار ما کفایت می‌کند.
و در آخر ایندکس FTS ایی را که پیشتر در مورد آن بحث کردیم، ایجاد خواهیم کرد:

```
-- Full-text index
CREATE FULLTEXT INDEX ON dbo.Documents
(
  docexcerpt Language 1033,
  doccontent TYPE COLUMN doctype
  Language 1033
  STATISTICAL_SEMANTICS
)
KEY INDEX PK_Documents
ON DocumentsFtCatalog
WITH STOPLIST = SQLStopList,
  SEARCH PROPERTY LIST = WordSearchPropertyList,
  CHANGE_TRACKING AUTO;
GO
```



در این تصویر محل یافتن اجزای مختلف Full text search را در management studio مشاهده می‌کنید.

یک نکته‌ی تکمیلی

برای زبان فارسی نیز یک سری stop words وجود دارند. لیست آن‌ها را از اینجا می‌توانید دریافت کنید:

stopwords.sql

متاسفانه زبان فارسی جزو زبان‌های پشتیبانی شده توسط FTS در SQL Server نیست (نه به این معنا که نمی‌توان با آن کار کرد؛ به این معنا که برای مثال دستورات صرفی زبان را ندارد) و به همین جهت از زبان انگلیسی در اینجا استفاده شده‌است.