DATA FROGS !

FRENCH SQL SERVER CONFERENCE

Microsoft SQL Server *Full-Text Search (FTS)* ou la **recherche textuelle**

Arian Papillon (MVP) – DATAFLY
Frédéric Brouard (SQLpro) – SQL SPOT

Recherche Textuelle - Principe

Faciliter la recherche d'éléments lexicaux dans de grand textes : mots, expressions, synonymes, formes fléchies…

L'usage du LIKE oblige à une double boucle :

- balayage des lignes
- parcours des caractères

…même si certains algorithmes (Boyer Moore) la rende un peu plus efficace sur le parcours des caractères…

Recherche Textuelle - Principe

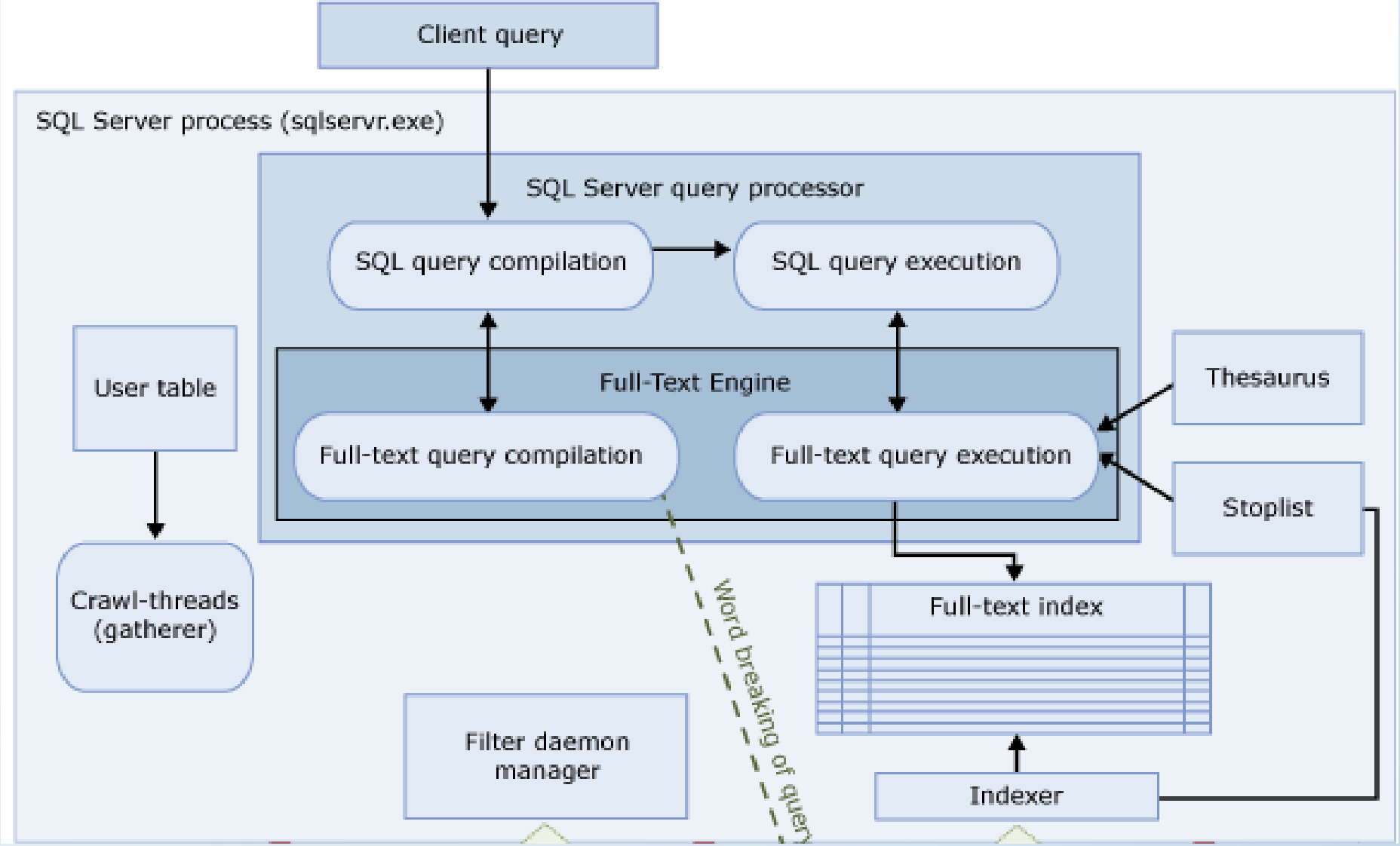
Norme SQL :

ISO/IEC 13249-2:2000
Information technology — Database languages — SQL multimedia and application packages — Part 2: Full-Text

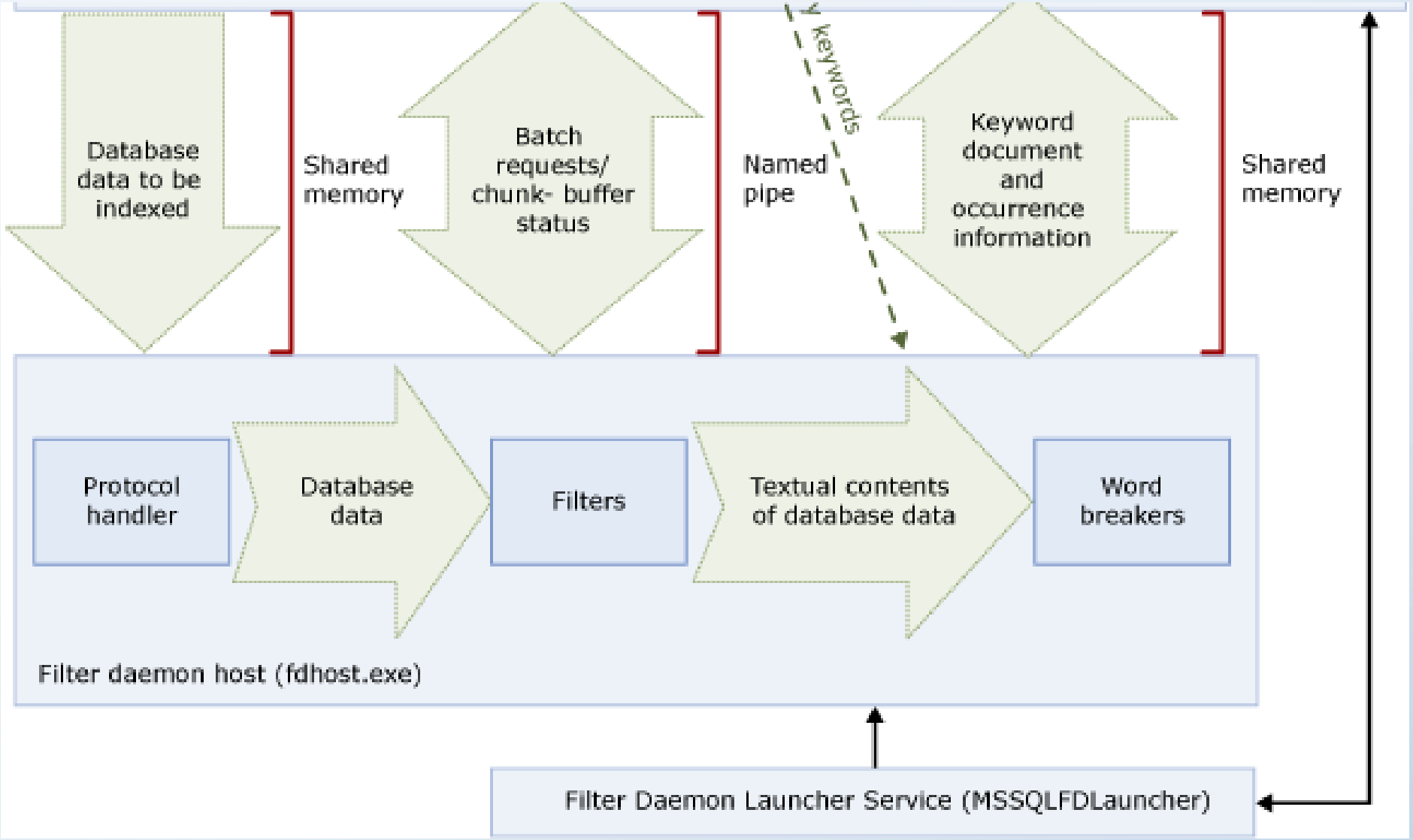
This standard has been revised by ISO/IEC 13249-2:2003

Introduit les fonctions CONTAINS, SCORE…

Recherche Textuelle – Principe dans SQL Server



Recherche Textuelle – Principe dans SQL Server



Recherche Textuelle – Mise en place

Création d'un espace de stockage

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

ATTENTION : la sensibilité aux accents est importante pour la plupart des langues latines excepté l'anglais (« maïs » <> « mais »)

Recherche Textuelle – Mise en place

Création d'un index textuel

```
CREATE FULLTEXT INDEX ON <table> (<liste_colonnes>)
   KEY INDEX <nom_index_unique_monocolonne>
   ON <catalogue / groupe_de_fichiers>
   WITH <liste__options>
GO
```

Options :

- CHANGE TRACKING : gère l'alimentation de l'index textuel
- STOPLIST : précise la liste des mots « noirs »
- SEARCH PROPERTY LIST : liste des propriétés des documents

Recherche Textuelle – Mise en place

Colonnes :

- Littérales : char, varchar, nchar, nvarchar, xml (JSON = NVARCHAR(max))
- Documents : VARBINARY(max) – directement ou indirectement (FILESTREAM, FileTable).

Mots noirs :

Liste de mots dépourvus de sémantique propre (articles, prépositions, pronoms, conjonctions de coordination…)

NOTA : SQL Server FTS indexe tous les mots sauf noirs, pourvu qu'ils soient formés de caractères, y compris, dates, ou encore de mots et chiffres et/ou caractères particuliers (tiret, espace, slash…). Ce n'est qu'à la recherche que les mots noirs sont éliminés

Recherche Textuelle – Mise en place

CHANGE TRACKING …

- AUTO : l'indexation est traitée au fil de l'eau
- MANUAL : à la demande, par journalisation ou en intégralité
- OFF [NO POPULATION] : désactivation de la journalisation

NOTA : l'alimentation de l'index textuel est toujours asynchrone pour des raisons de performance…

STOPLIST : précise le liste des mots « noirs »

- soit celle par défaut (une par langue - SYSTEM)
- soit créée de toute pièce (CREATE/ALTER/DROP FULLTEXT STOPLIST)
- soit aucune (OFF)

Recherche Textuelle – Recherches

4 fonctions :

Fonction	Description
CONTAINS	Recherches précises (norme SQL)
FREETEXT	Recherches « vagues »
CONTAINSTABLE	Recherches précises pondérées
FREETEXTTABLE	Recherches « vagues » pondérées

CONTAINS et FREETEXT sont des fonctions scalaires
CONTAINSTABLE et FREETEXTTABLE sont des fonctions table

Recherche Textuelle – Recherches

CONTAINS … :

Prédicat	Type de recherche
« mot »	Un mot précis
« mot1-mot2… »	Une expression précise
« mo* »	Mot commençant par
NEAR(…)	Termes proches, limités en distance, en ordre ou pas
INFLECTIONAL	Forme fléchie
THESAURUS	Synonyme, acronymes

Ces membres de prédicats peuvent être assemblés à l'aide des connecteurs logique AND, OR et NOT.

Recherche Textuelle – Recherches

CONTAINS et CONTAINSTABLE :

- repose sur les principes vus ci-avant

FREETEXT et FREETEXTTABLE

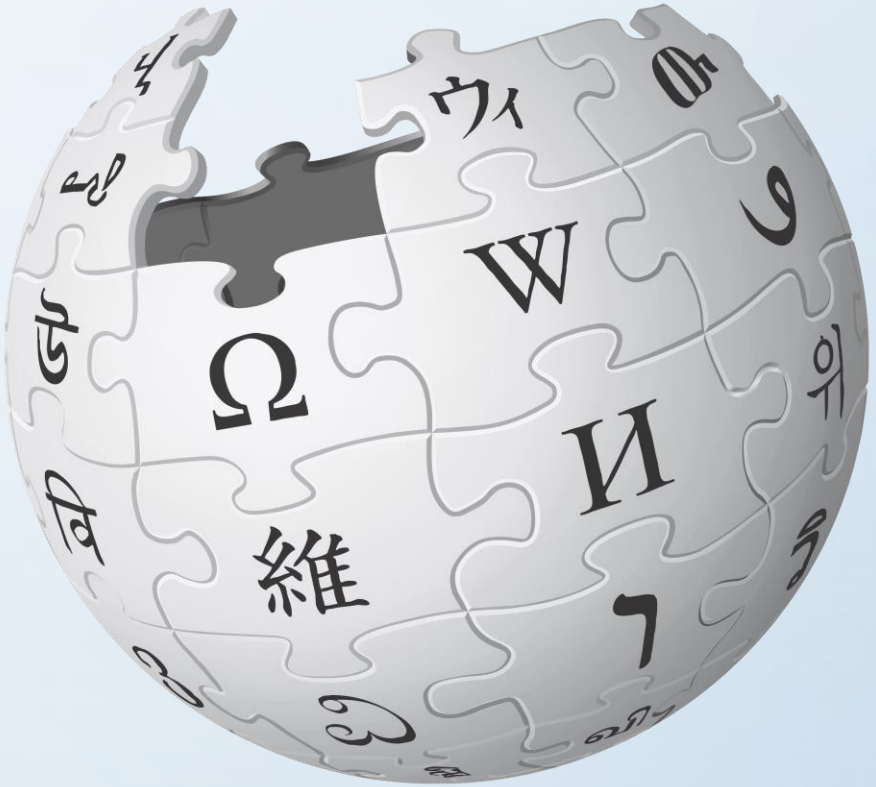
- nécessite uniquement une liste de mots (recherche « à la google »)

CONTAINSTABLE et FREETEXTTABLE :

- permettent des recherches pondérées et offrent des résultats indicés (équivalent de la fonction SCORE de la norme SQL)

Démo

Recherche Full-Text



Recherche Textuelle – Documents électroniques

Possible via VARBINARY(max)

- directement
- indirectement via FILESTREAM ou FileTable

Nécessite l'utilisation de « iFilter » :

- DLL extrayant le texte de documents et certaines propriétés (méta tags);
- Une cinquantaine en standard (formats libre + MS Office)

Propriétés des documents :

- Titre, auteur…
- Liste obtenable par FILTDUMP.exe
- Interrogation via PROPERTY ({ nom_colonne}, ' nom_propriété ')

Démo

Recherche dans des documents

Recherche Textuelle – Recherche sémantique

Recherche sémantique

- Porte sur la signification (ce qui a du sens) et non plus les mots…
- Basée sur les statistiques d'apparition des mots
- Nécessite une base de données (semanticdb – une par langue) pourvue par Microsoft, ayant ces statistiques (analyse lexicales de millions de textes de toute nature).

A partir d'un texte « source » :

- Contenu dans la table
- Soit un texte, soit le contenu textuel d'un document

Recherche Textuelle – Recherche sémantique

Trois fonctions table :

SEMANTICKEYPHRASETABLE : retourne une table contenant les expressions clés associées aux colonnes de la table spécifiée.

SEMANTICSIMILARITYTABLE : retourne une table de zéro, une ou plusieurs lignes pour les documents dont le contenu dans la colonne spécifiée est sémantiquement similaire à un document spécifié.

SEMANTICSIMILARITYDETAILSTABLE : retourne une table d'expressions clés communes à travers deux documents (le document source et un document mis en correspondance) dont le contenu est sémantiquement similaire.

Démo

Recherche sémantique

Recherche Textuelle – Pour aller plus loin

DIFFICULTÉS :

Polysémie :

- Des mots ayant une même graphie mais des sens différents (avocat, glace, côte…) ou pire… contraires (louer, apprendre, plus…)
- Provoque des faux positifs…

Synonymes :

- Doit être strictement contrôlé par le thésaurus. Une synonymie est propre à un univers de mots. Exemple : « four »…
  - fourneau, foyer, cuisinière (cuisine)
  - bide, échec, insuccès (show-business)

Recherche Textuelle – Pour aller plus loin

MANQUES :

Mots se terminant par (LIKE '%mot') :

- contournement facile par ajout d'une colonne inversant les littéraux.
- pour les doc électroniques : récupérer la liste des mots indexés via :

`sys.dm_fts_index_keywords`

… créer une table de ces mots inversés puis utiliser

`sys.dm_fts_index_keywords_position_by_document`

… pour les liens mots/lignes + position

Recherche Textuelle – Pour aller plus loin

MANQUES :

Mots contenant (LIKE '%mot%') :

- utiliser les fonctions:

`sys.dm_fts_index_keywords`
`sys.dm_fts_index_keywords_position_by_document`

… pour récupérer les mots et les liens mots/lignes + position
Puis créer une table des mots / ligne / Position et une table pour un index rotatif ou une solution par trigramme.

https://blog.developpez.com/sqlpro/p13123
https://sqlperformance.com/2017/09/sql-performance/sql-server-trigram-wildcard-search

Recherche Textuelle – Pour aller plus loin

MANQUES :

Mots mal orthographié (« guithare ») :

- Utiliser l'algorithme de KNUTH : si l'on coupe en deux un mot mal orthographié en son milieu, alors l'erreur est à droite ou à gauche…
- Il ne reste plus qu'à comparer les demi-mots droite et gauche à l'ensemble des mots

Affinement de la comparaison par scoring avec algorithmes de HAMMING, LEVENSTHEIN ou inférence basique :

https://sqlpro.developpez.com/cours/soundex/
https://blog.developpez.com/sqlpro/p8243/

Recherche Textuelle – Pour aller plus loin

MANQUES :

Découpage des documents par chapitre, page, phrase…

- Prévu par la norme SQL
- Aucun système ne sait comment faire (quel caractère non imprimable ?)
- La notion de page diffère à l'impression, au cadrage , au dimensionnement…

Inconvénient : la recherche textuelle vous indique le document, mais pas la position dans le document.
Éventuellement calculer un indice positionnel :

- 100.0 * position du mot dans le doc / nombre de mot du doc (en %)

Recherche Textuelle – Pour aller plus loin

DOCUMENTATION

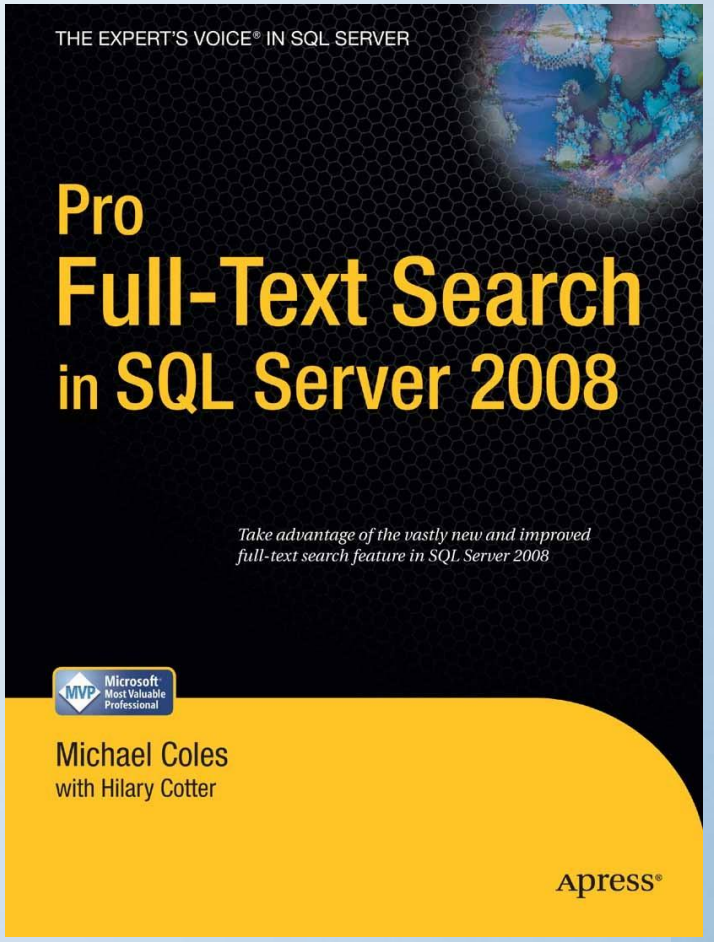
Livre : Pro Full-Text Search in SQL Server 2008

WEB :
https://www.sqlshack.com/hands-full-text-search-sql-server/

https://www.red-gate.com/simple-talk/databases/sql-server/learn/understanding-full-text-indexing-in-sql-server/

https://www.linkedin.com/pulse/sql-server-like-vs-fulltext-search-comparison-bhanu-akaveeti/

https://blog.developpez.com/sqlpro/p9344




partie
examiner
technique
faire
comme
intégral
consiste
correspondent
préférée
parcours
position
méthode
appelée
document
sauf
séquentiel
entrée
faite
indexation
avoir
Alta
être
1
plein
très
base
Vista
1970
presque
Bien
tels
petits
recherche
mots
simple
mot
crée
telle
aide
on
QUESTIONS
2
requête
ceux
Il
Des
L
autres
trop
sont
La
texte
Web
liste
Les
corpus
sites
utiles
exacte
sans
puisse
plus
documents
chaque
données
outils
libre
moteurs
récupérer
Pour
aussi
index
scanner
moteur
techniques
années
électronique
fréquents
toutes
recherches