Strategic Sponsor

Microsoft

Gold Sponsors

PYRAMID ANALYTICS

KRAJOWY REJESTR DŁUGÓW www.krd.pl

devart Database Interaction Solutions

Silver Sponsors

Profisee®

SQL EXPERT.pl

CREDIT SUISSE

Technical Partners

Microsoft

PTI POLSKIE TOWARZYSTWO INFORMATYCZNE ODDZIAŁ GÓRNOŚLĄSKI

Academic Partners

Wyższa Szkoła Zarządzania i Bankowości w Krakowie

Wyższe Szkoły Bankowe

Media Partners

GEEKCLUB.pl wss.pl

ERP24.pl

4programmers.net

IT professional

IN FACT PR

itnews24

virtual-it.pl

IT WORKS POLSKA

VirtualStudy

MAGAZYN IT PL

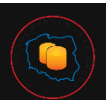
Made in Wro Do IT here!

ERP-view.pl Portal rozwiązań IT w biznesie

Wyszukiwanie pełnotekstowe (Full-Text Search) w SQL Server

Kamil Nowiński
PLSSUG Wrocław
kamil.nowinski@plssug.org.pl

Tomasz Libera
PLSSUG Kraków
tomasz.libera@plssug.org.pl



Kamil Nowiński

- Senior SQL/BI Developer w AlternativeNetworks (UK)
- Programista > 20 lat
  (VB6, VB.NET, C#, .NET Framework)
- Ponad 10-letnie doświadczenie jako DEV/DBA
- Członek komisji rewizyjnej PLSSUG,
- Co-Leader PLSSUG Wrocław
- Certyfikaty SQL Server:
  MCITP, MCP, MCTS, MCSA,
  MCSE Data Platform
- Zainteresowania:
  - Bieganie – obecnie trening do półmaratonu,
  - Fotografia cyfrowa
    (Nikon D-90, Adobe Lightroom)

Microsoft CERTIFIED IT Professional

MCSE Data Platform

Microsoft CERTIFIED Solutions Associate SQL Server 2012

Tomasz Libera

- DB Developer w WSZiB w Krakowie
- Lider PLSSUG Kraków
- Certyfikaty:
  - MCT
  - MCSE Data Platform
  - MCITP-DBA, MCITP-DD
- Zainteresowania:
  - Pasjonat kolarstwa górskiego i maratonów MTB

Agenda

- Wprowadzenie
  - LIKE to za mało
  - Możliwości, Historia
  - Architektura, Komponenty FTS
- Zapytania - CONTAINS & FREETEXT
- Indeksy FTS – budowa, tworzenie
- Wyrazy szumy, słownik synonimów, wyszukiwanie semantyczne
- Wyszukiwarka SQLoogle!

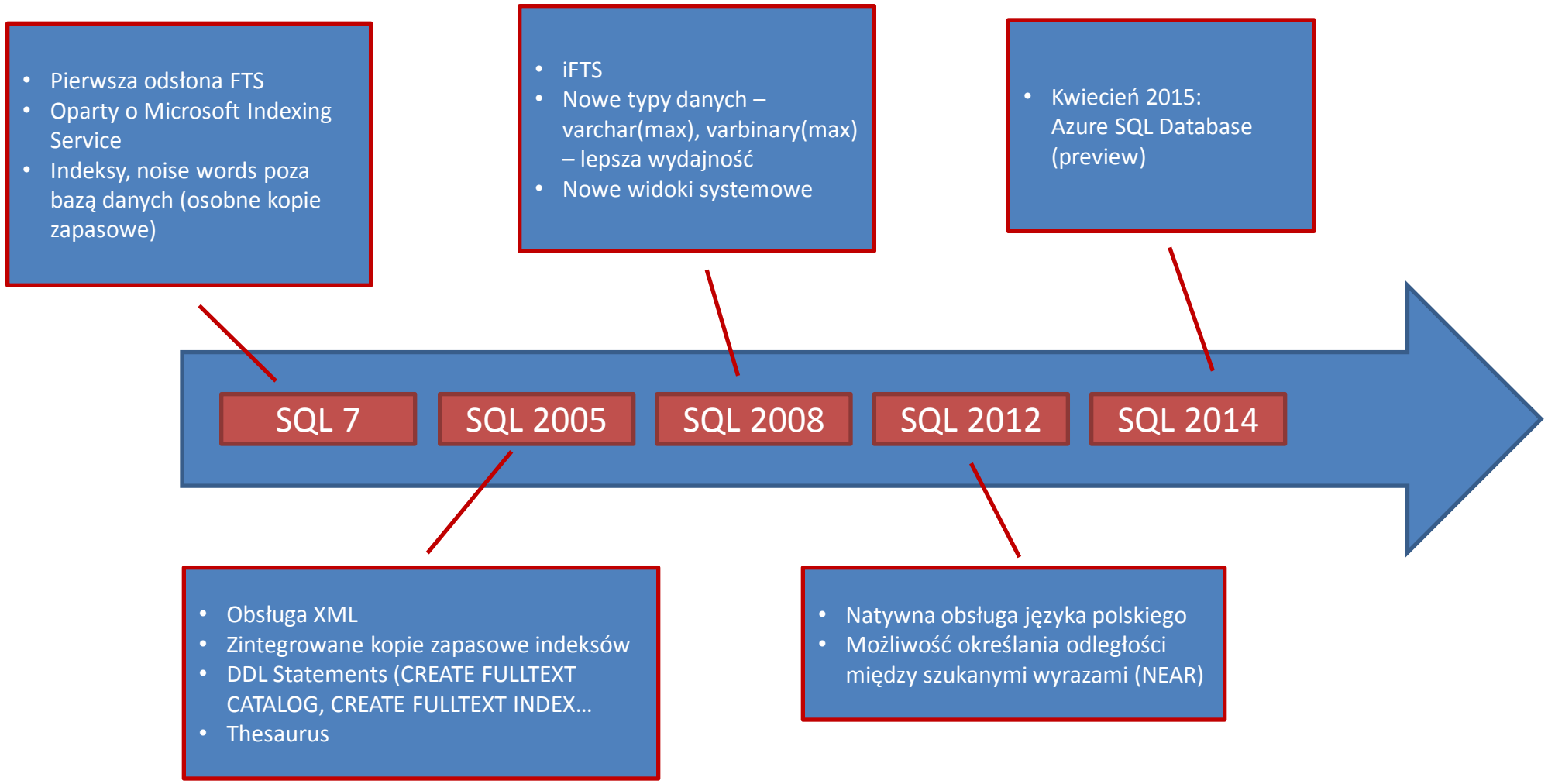
DEMO #1

Like to za mało!

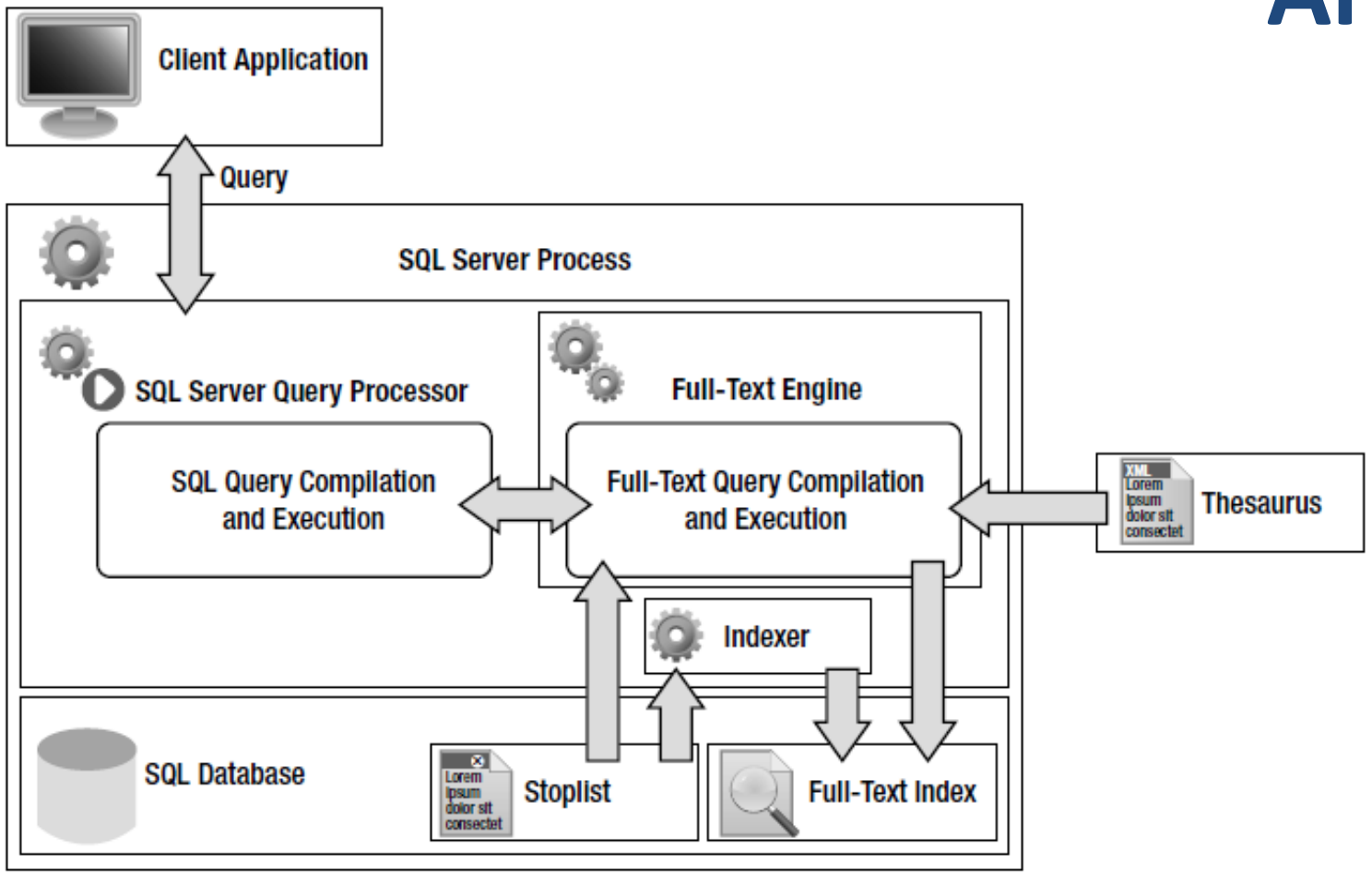
Możliwości

- Wyszukiwanie danych tesktowych zapisanych w
  - kolumnach tekstowych (varchar, nvarchar)
  - danych binarnych w obsługiwanych formatach (txt, doc, *docx*, *pdf*...)
- Dostępne już w bezpłatnej edycji SQL Server Express
- Wyrażenia proste – jedno bądź wiele słów
- Poszukiwanie różnych form gramatycznych
- Wyrazy bliskoznaczne, synonimy słów
- Sąsiedztwo wyrażeń (słów lub fraz)
- Ważenie wyrażeń
- Pozycjonowanie (rangowanie) wyników
- Wyszukiwanie semantyczne

Historia


Pierwsza odsłona FTS
Oparty o Microsoft Indexing Service
Indeksy, noise words poza bazą danych (osobne kopie zapasowe)
iFTS
Nowe typy danych – varchar(max), varbinary(max) – lepsza wydajność
Nowe widoki systemowe
Kwiecień 2015: Azure SQL Database (preview)
SQL 7
SQL 2005
SQL 2008
SQL 2012
SQL 2014
Obsługa XML
Zintegrowane kopie zapasowe indeksów
DDL Statements (CREATE FULLTEXT CATALOG, CREATE FULLTEXT INDEX…
Thesaurus
Natywna obsługa języka polskiego
Możliwość określania odległości między szukanymi wyrazami (NEAR)

Architektura



Pro Full-Text Search in SQL Server 2008, Michael Coles, Apress

Zapytania FTS

- Operatory porównania wykorzystywane w części WHERE zapytania
  - CONTAINS
  - FREETEXT
- Funkcje tabelaryczne (w części FROM), umożliwiają rangowanie
  - CONTAINSTABLE
  - FREETEXTABLE

DEMO #2

Zapytania

Predykaty CONTAINS & FREETEXT

	CONTAINS	FREETEXT
Formy fleksyjne wyrazów	Na żądanie FORMSOF(INFLECTIONAL, wyraz)	Zawsze
Thesaurus (w tym synonimy)	Na żądanie FORMSOF(THESAURUS, wyraz)	Zawsze
Wagi dla wyrazów	TAK	NIE
Operatory logiczne	TAK	NIE
Wyrazy blisko siebie	TAK	NIE
Przedrostki wyrazów (prefix)	TAK	NIE
Zapytanie	Złożone, większa kontrola	Dużo prostsze, mniejsza kontrola
Razem rezultaty	Mniejsza liczba wyników Dokładniejsze	Większa liczba wyników Mniej precyzyjne

Wykonywanie zapytania pełnotekstowego

Silnik Full-Text Search:

1) Wykonuje dzielenie wyrazów (*word breaker*) frazy
2) Generuje formy fleksyjne (*steamer*)
3) Identyfikuje listę rozszerzeń i zastąpień (*thesaurus*)
4) Znajduje wszystkie powyższe wyrazy

Budowa indeksu

display_term	DocID	occurrence	special_term
dom	1	1	Exact Match
to	1	2	Noise Word
nie	1	3	Noise Word
tylko	1	4	Exact Match
budynek	1	5	Exact Match
dom	1	14	Exact Match
to	1	15	Noise Word
mama	1	16	Exact Match
tata	1	17	Exact Match
i	1	18	Noise Word
ja	1	19	Noise Word
to	2	1	Noise Word
moja	2	2	Noise Word
babcia	2	3	Exact Match
i	2	4	Noise Word
maly	2	5	Exact Match
brat	2	6	Exact Match

http://pliki.naszelementarz.men.gov.pl/elementarz/naszelementarz.pdf

```
SELECT * FROM sys.dm_fts_parser ('"Dom to nie tylko budynek. Dom to mama, tata i ja."', 1045, 0, 0)
SELECT * FROM sys.dm_fts_parser ('"To moja Babcia i mały brat."', 1045, 0, 0);
```

Aktualizacja (przebudowa indeksu)

Asynchroniczna:

- Pełna

```
ALTER FULLTEXT INDEX ON NewsPL
START FULL POPULATION
```

- Przyrostowa – tylko wiersze zmodyfikowane od ostatniego wypełniania (wymaga kolumny RowVersion w tabeli)

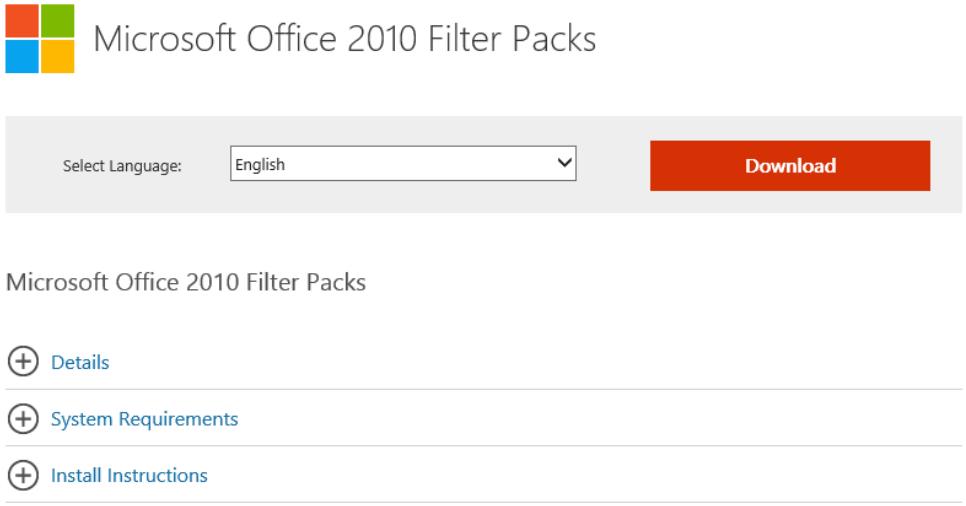
```
ALTER FULLTEXT INDEX ON NewsPL
START INCREMENTAL POPULATION
```

- Tylko wiersze zmodyfikowane od ostatniego wypełniania, wymaga CHANGE_TRACKING (nie wymaga RowVersion)

```
ALTER FULLTEXT INDEX ON NewsPL
START UPDATE POPULATION
```

Obsługa plików docx

- MS Office 2010 Filter Packs
- sp_fulltext_service 'load_os_resources', 1
- Restart instancji



Supported file types

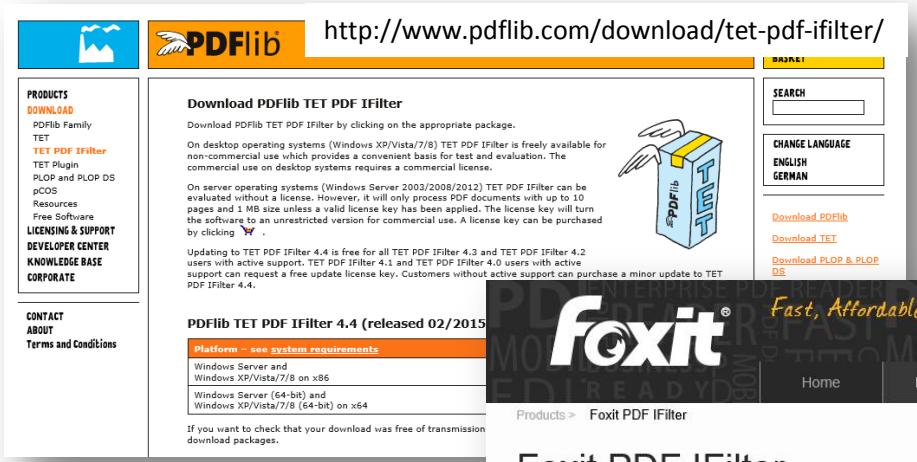
The following file types are supported by the Office 2007 Filter Pack:

.docx	.docm	.pptx	.pptm	.xlsx
.xlsm	.xlsb	.zip	.one	.vdx
.vsd	.vss	.vst	.vsx	.vtx

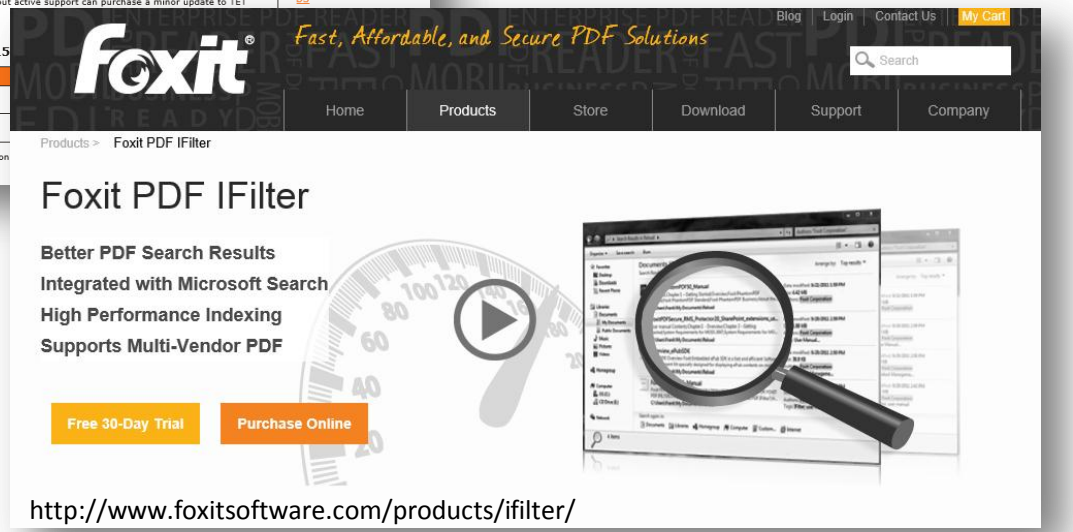
The following file types are supported by the Office 2010 Filter Pack:

.docx	.docm	.dotx	.pptx	.pptm	.xlsx
.xlsm	.zip	.html	.doc	.xls	.ppt
.dot	.vss	.vsd	.vst	.vdx	.vsx
.vtx	.vdw	.one	.odt	.odp	.ods
.msg	.pub				

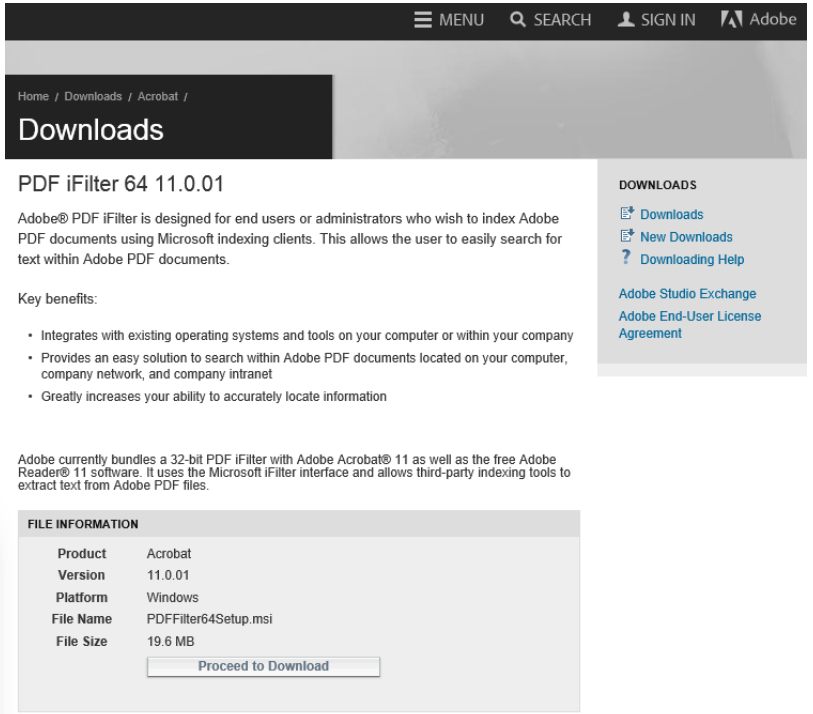
Obsługa plików PDF



http://www.pdflib.com/download/tet-pdf-ifilter/



http://www.foxitsoftware.com/products/ifilter/



http://www.adobe.com/support/downloads/detail.jsp?ftpID=5542

DEMO #3

katalogi, Indeksy

Stoplists & stopwords

- *Stoplist*-a może zawierać wiele *stopword's*
- Pozwala zapobiegać indeksowaniu (a dokładniej wyszukiwaniu)
- Przykłady:
  - łączniki: i, lub, oraz,
  - Słowa często występujące w danej dziedzinie (np. SQL)

- Jak sprawdzić:

```
SELECT * FROM sys.fulltext_stoplists;
SELECT * FROM sys.fulltext_stopwords;
```

Synonimy i plik tezaurus (thesaurus)

- Przechowywane na dysku MSSQL\FTDATA tsXXX.xml - (XXX - ENU,PLN)
- tsglobal - globalny tezaurus- wykorzystywany dodatkowo, bez względu na jaki język
  - jeśli poszukiwany wyraz znajduje się w globalnym i lokalnym (specyficznym dla języka) - pierwszeństwo ma lokalny
- Edycja pliku pozwala na konfigurację:
  - Diacritics_sensitive
  - **Expansion** (autor, pisarz)
    wyszukiwane są również słowa bliskoznaczne
  - **Replacement** (Win 2k8 -> Windows 2008)
    wyszukiwane słowo jest zastępowane słowem bliskoznacznym

- Po edycji pliku musi zostać on załadowany przez SQL Server:

```
EXEC sys.sp_fulltext_load_thesaurus_file 1033;
```

Wyszukiwanie semantyczne

- Pozwala na głębsze wniknięcie do dokumentów:
  - Indeksowanie statystyczne zależnych fraz kluczowych
- Wykorzystanie fraz kluczowych służy:
  - Wyszukiwaniu dokumentów podobnych lub powiązanych
- *Semantic search* rozszerza możliwości wyszukiwania pełnotekstowego
- Wymagania:
  - Zainstalowana baza danych **Semantic Language Statistics**
    Dostępna na płycie z instalatorem SQL Server w folderze:
    \x64\Setup\SemanticLanguageDatabase.msi

DEMO #4

stoplist, thesaurus, DMV

DEMO #5

wyszukiwarka SQLoogle

http://youtu.be/LY-X1LaPFp4

Q&A

ANY
QUESTIONS

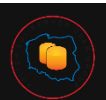
DZIĘKUJEMY ZA UWAGĘ!

Kamil Nowiński
PLSSUG Wrocław
kamil.nowinski@plssug.org.pl
@NowinskiK

Tomasz Libera
PLSSUG Kraków
tomasz.libera@plssug.org.pl
@tomasz_libera

PLIKI

http://1drv.ms/1AsZoID



Strategic Sponsor

Microsoft

Gold Sponsors

PYRAMID ANALYTICS

KRAJOWY REJESTR DŁUGÓW www.krd.pl

devart Database Interaction Solutions

Silver Sponsors

Profisee®

SQL EXPERT.pl

CREDIT SUISSE

Technical Partners

Microsoft

PTI POLSKIE TOWARZYSTWO INFORMATYCZNE ODDZIAŁ GÓRNOŚLĄSKI

Academic Partners

Wyższa Szkoła Zarządzania i Bankowości w Krakowie

Wyższe Szkoły Bankowe

Media Partners

SPECJALISTA IT PROGRAMISTA DESIGNER GEEKCLUB.pl wss.pl

ERP24.pl

4programmers.net

IT professional

IN FACT PR

itnews24

Vir virtual-it.pl

iT NEWS POLSKA

VirtualStudy

MAGAZYN IT PL

Made in Wro Do IT here!

ERP-view.pl Portal rozwiązań IT w biznesie