Sperasoft

Your game dev partner

# Business Intelligence

## Semantic Search in SQL Server 2012

# What is Semantic Search

- Semantic search seeks to improve search accuracy by understanding searcher intent and **the contextual meaning of terms** as they appear in the searchable dataspace.

# Semantic Search in SQL Server 2012

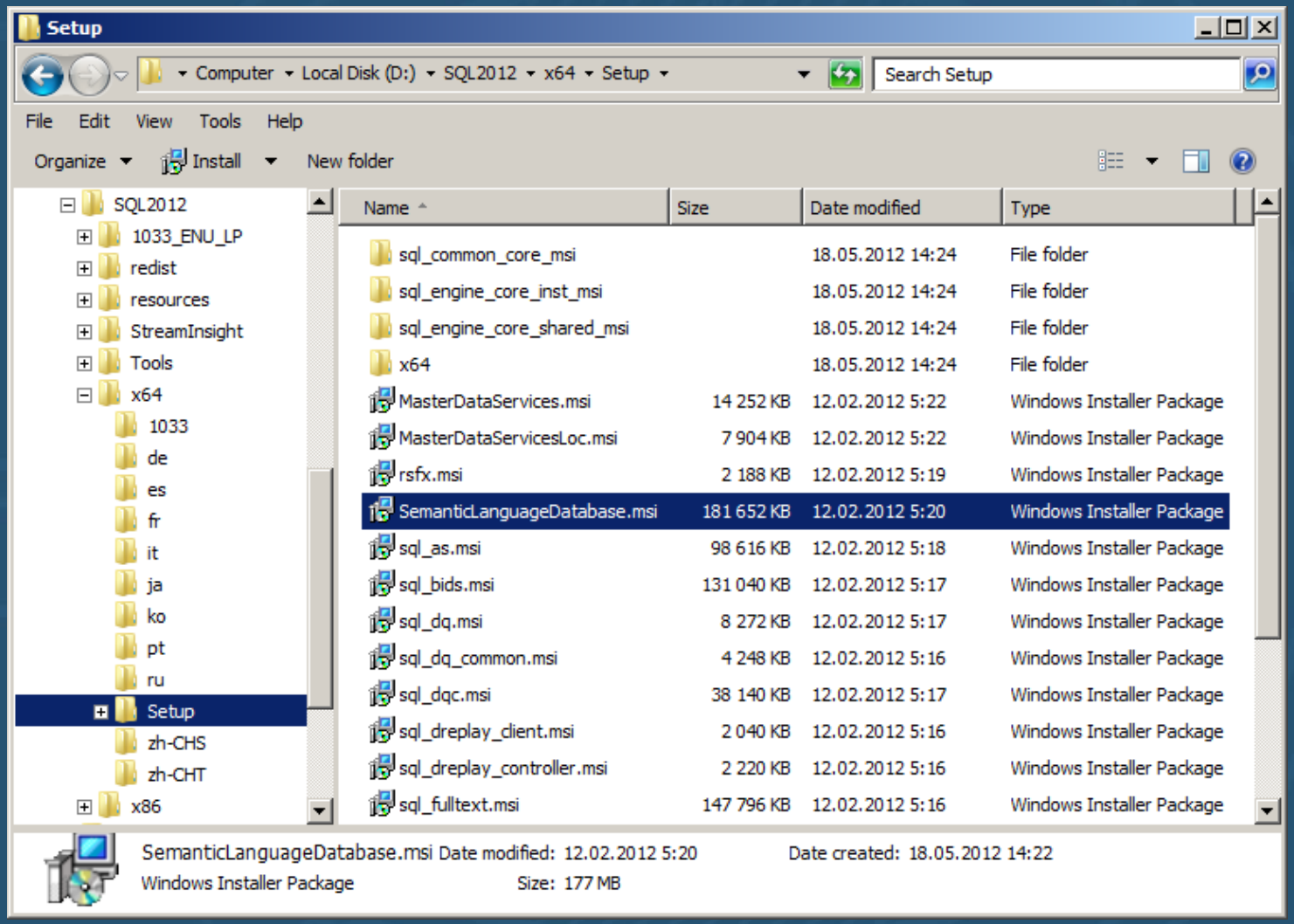
- Built on top of Full-Text Search
- Requires predefined external Database
- That database should be attached to SQL Server Instance
- Semantic Search should be configured to use that Database

# Supported in SQL Server Editions

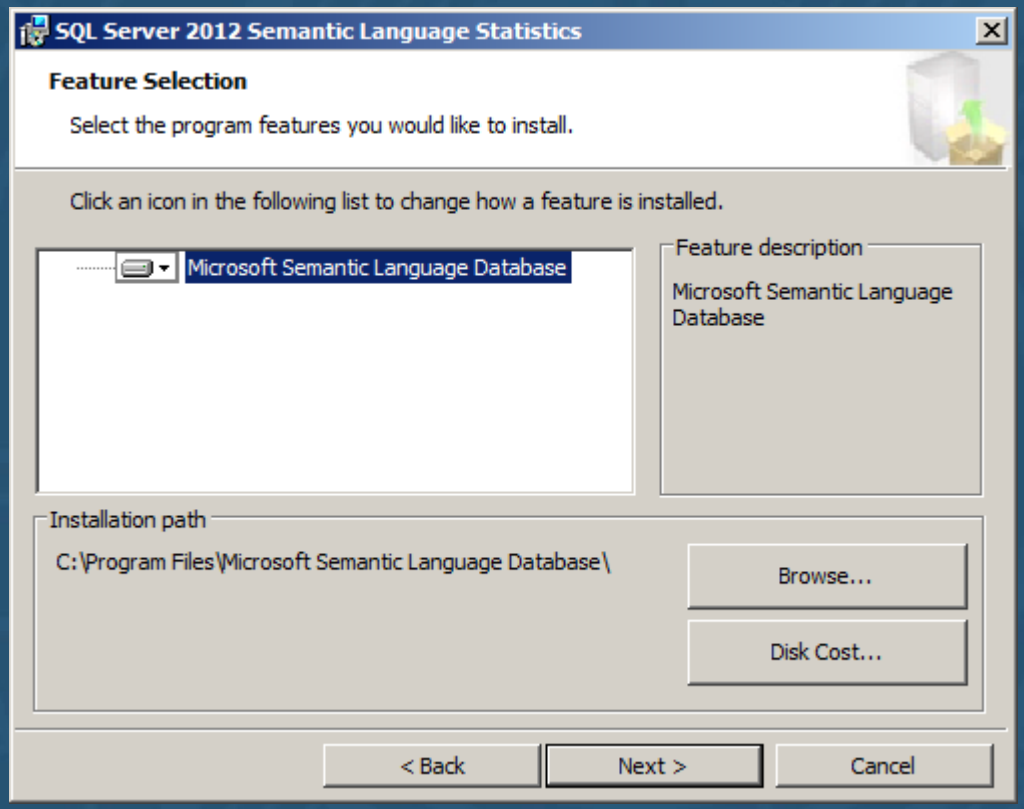
- Exists in all Commercial editions of SQL Server 2012
- Also in SQL Server 2012 Express Advanced Services Edition

# Semantic Search Installation 1/3

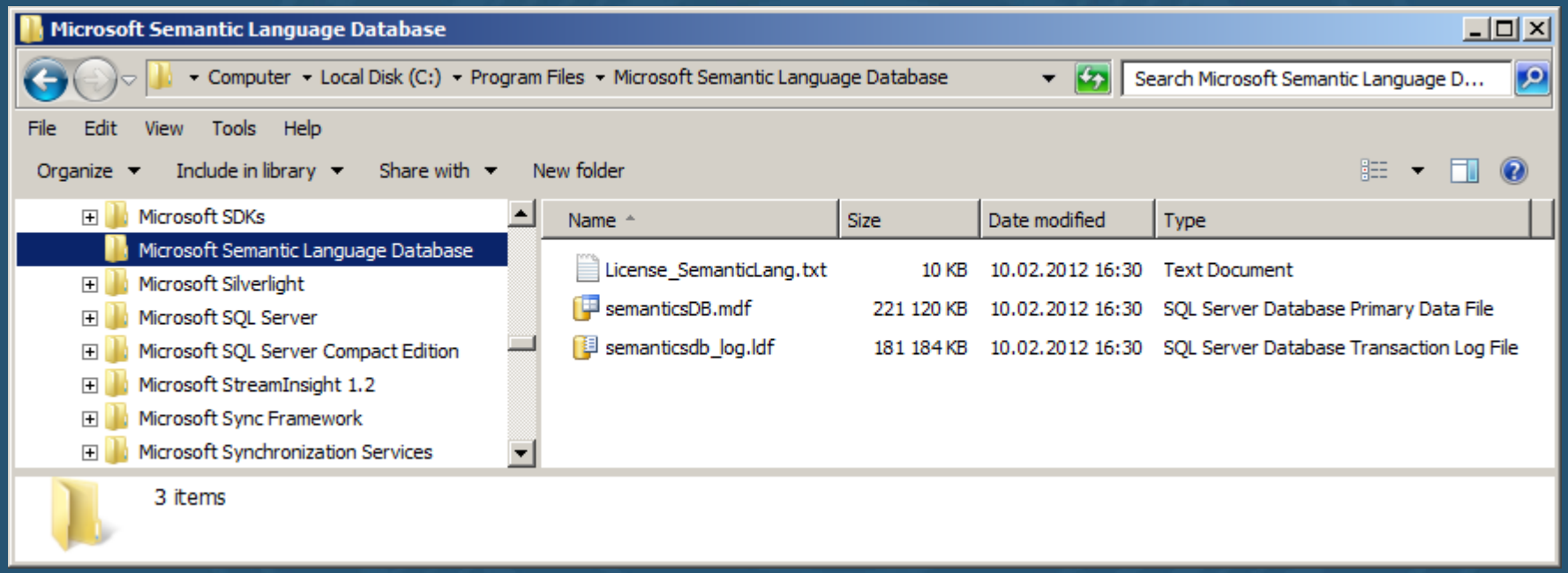
Sperasoft
Your game dev partner



# Semantic Search Installation 2/3



# Semantic Search Installation 3/3



# Attach Semantics DB

```
-- do not use sp_attach_db stored procedure
-- it is obsolete

CREATE DATABASE SemanticsDB
  ON (FILENAME = N'C:\Program Files\Microsoft Semantic Language Database\semanticsDB.mdf')
  LOG ON (FILENAME = 'C:\Program Files\Microsoft Semantic Language Database\semanticsdb_log.ldf')
  FOR ATTACH;
GO
```

# Register Semantics DB

```
-- Register Semantics Languages Database
-- required once

EXEC
sp_fulltext_semantic_register_language_statisti
cs_db @dbname = N'SemanticsDB';
GO
```

# Verify Registration

-- Verify the registration is succeeded

SELECT * FROM sys.fulltext_semantic_language_statistics_database;
GO

```
19
20  -- Verify the registration is succeeded
21  SELECT * FROM sys.fulltext_semantic_language_statistics_database;
22  GO
```

100 %

Results | Messages

| | database_id | register_date | registered_by | version |
|---|---|---|---|---|
| 1 | 18 | 2012-05-18 16:03:37.960 | 1 | 11.0.1153.1.1 |

# Supported Languages

-- Check available languages for statistical semantic extraction

SELECT * FROM sys.fulltext_semantic_languages;
GO

```
15  -- Check available languages for statistical semantic extraction
16  SELECT * FROM sys.fulltext_semantic_languages;
17  GO
```

100 %

Results | Messages

| | lcid | name |
|---|---|---|
| 1 | 1028 | Traditional Chinese |
| 2 | 1031 | German |
| 3 | 1033 | English |
| 4 | 1036 | French |
| 5 | 1040 | Italian |
| 6 | 1046 | Brazilian |
| 7 | 1049 | Russian |
| 8 | 1053 | Swedish |
| 9 | 2052 | Simplified Chinese |
| 10 | 2057 | British English |
| 11 | 2070 | Portuguese |
| 12 | 3076 | Chinese (Hong Kong SAR, PRC) |
| 13 | 3082 | Spanish |
| 14 | 4100 | Chinese (Singapore) |
| 15 | 5124 | Chinese (Macau SAR) |

# How to Enable On Table

Sperasoft
Your game dev partner

Demo

# Restart Processes

```
-- Reload filters (iFilter) and restart fulltext
-- host process if needed

EXEC sp_fulltext_service 'load_os_resources', 1;
EXEC sp_fulltext_service 'restart_all_fdhosts';
GO
```

# Full-Text Search

- Supports character-based columns:
  1. char
  2. varchar
  3. nchar
  4. nvarchar
  5. text
  6. ntext
  7. image
  8. xml
  9. varbinary (max)
  10. FileStream

# Full-Text Queries Specifics

- Full-text queries are not case-sensitive searching for "Aluminum" or "aluminum" returns the same results

- Transact-SQL predicates:
  - CONTAINS
  - FREETEXT

- Transact-SQL functions:
  - CONTAINSTABLE
  - FREETEXTTABLE

# File types supported by iFilters

SELECT * FROM sys.fulltext_document_types;

```
SELECT * FROM sys.fulltext_document_types;
```

100 %

Results | Messages

| | document_type | class_id | path | version | manufacturer |
|---|---|---|---|---|---|
| 1 | .ascx | E0CA5340-4534-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 2 | .asm | C7310720-AC80-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 3 | .asp | E0CA5340-4534-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 4 | .aspx | E0CA5340-4534-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 5 | .bat | C7310720-AC80-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 6 | .c | C7310720-AC80-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 7 | .cmd | C7310720-AC80-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |
| 8 | .cpp | C7310720-AC80-11... | C:\Program Files\Micro... | 12.0.6828.0 | Microsoft Corporation |

# Semantic Search Functions

Three Tabular Functions:

- SemanticKeyPhraseTable - returns the statistically significant phrases in each document
- SemanticSimilarityTable – returns documents or rows that are similar or related, based on the key phrases in each document
- SemanticSimilarityDetailsTable – returns the key phrases that explain why two documents were identified as similar

# Full-Text Catalog Items Count

```
-- select Full-Text Catalog items count
SELECT FulltextCatalogProperty
   ('FullTextCatalog', 'itemcount');
GO
```

# Full-Text Catalog Population Status

```
-- check Population progress
SELECT fulltextcatalogproperty('FullTextCatalog', 'populatestatus');
GO
```

- 0 = Idle
- 1 = Full population in progress
- 2 = Paused
- 3 = Throttled
- 4 = Recovering
- 5 = Shutdown
- 6 = Incremental population in progress
- 7 = Building index
- 8 = Disk is full. Paused.
- 9 = Change tracking

# Get all Key Phrases

```
-- Get all key phrases in the entire corpus

SELECT
  K.score, K.keyphrase, COUNT(D.stream_id) AS Occurrences
  FROM SemanticKeyPhraseTable
    (dbo.Documents, (name, file_stream)) AS K
  INNER JOIN dbo.Documents AS D
    ON D.path_locator = K.document_key
  GROUP BY K.score, K.keyphrase
  ORDER BY K.score DESC, K.keyphrase ASC;
GO
```

# Find Documents by Key phrase

```
-- Find documents by keyphrase – ‘sql’ in the case below

SELECT
  K.score, K.keyphrase,
  D.stream_id, D.name, D.file_type, D.cached_file_size,
  D.creation_time, D.last_write_time, D.last_access_time
FROM dbo.Documents D
INNER JOIN semantickeyphrasetable (
  dbo.Documents,
  (name, file_stream)
) AS K
ON D.path_locator = K.document_key
WHERE K.keyphrase = N'sql'
ORDER BY K.score DESC;
```

# Find Similar Documents

```
-- find similar documents
DECLARE @Title NVARCHAR(1000) = (SELECT'Gurevich Vladimir.docx');
DECLARE @DocID HIERARCHYID =
(SELECT path_locator FROM dbo.Documents WHERE name = @Title);

SELECT
  @Title AS source_title, D.name AS matched_title,
  D.stream_id, K.score
  FROM SemanticSimilarityTable(dbo.Documents, *, @DocID) AS K
  INNER JOIN dbo.Documents AS D
    ON D.path_locator = K.matched_document_key
  ORDER BY K.score DESC;
GO
```

# Why 2 Documents Are Similar

```
-- find out Key Phrases that make two documents match

DECLARE @SourceTitle NVARCHAR(1000) = (SELECT ‘source.docx');
DECLARE @MatchedTitle NVARCHAR(1000) = (SELECT ‘target.docx');

DECLARE @SourceDocID HIERARCHYID =
   (SELECT path_locator FROM dbo.Documents WHERE name = @SourceTitle);
DECLARE @MatchedDocID HIERARCHYID =
   (SELECT path_locator FROM dbo.Documents WHERE name = @MatchedTitle);

SELECT
  K.keyphrase, K.score, @SourceTitle AS source_title, @MatchedTitle AS matched_title
  FROM SemanticSimilarityDetailsTable(dbo.Documents, file_stream, @SourceDocID,
file_stream, @MatchedDocID) AS K
  ORDER BY K.score DESC;
GO
```

# Full-Text Search NEAR Operator 1/2

- The generic NEAR operator is deprecated in SQLServer2012
- It is a new operator and not an extension of the existing NEAR operator

- Lets to query with 2 optional requirements that you could not previously specify
  1. The maximum gap between the search terms
  2. The order of the search terms - for example, “John” must appear before “Smith”

- Stopwords or noise words are included in the gap count.

CONTAINSTABLE(Documents, Content, ‘NEAR((John, Smith), 4, TRUE)’);

# Full-Text Search NEAR Operator 2/2

- -- get documents that contain keywords "sql" and "server" nearby
- SELECT D.name, file_stream.GetFileNamespacePath() AS relative_path
- FROM dbo.Documents D
- WHERE CONTAINS(file_stream, 'NEAR(("sql", "server"), 1, FALSE)');
- GO

# Full-Text Search in Documents

```
-- get documents that contain keywords "sql" and "server" nearby

SELECT D.name, file_stream.GetFileNamespacePath() AS relative_path
  FROM dbo.Documents D
  WHERE CONTAINS
(file_stream, 'NEAR(("sql", "server"), 1, FALSE)');
GO
```

# Problems

- Full Text Catalog depend on language selected