

Searching Unstructured Text in DB2 for z/OS: When LIKE Isn't Good Enough

DB2 Information Management Software

J. Mark Wilson
IBM Americas S&D
DB2 for z/OS
mwilson2@us.ibm.com

ON DEMAND BUSINESS™

Agenda


- What is unstructured text?
- Previous solution limitations and requirements
- Architecture
- Searching
- Administration

Structured data

- Most traditional relational data is structured:
 - Customer first, middle, last names
 - Phone numbers
 - City, state, and zip code
 - SSN
 - Patient account number
 - . . .

3

Example of unstructured text: US Customs data

HORSES AND ASSES, PUREBRED BREEDING, LIVE	
HORSES, LIVE, EXCEPT PUREBRED BREEDING	
ASSES, MULES, AND HINNIES, LIVE, NESOI	
BOVINES, PUREBRED BREEDING, DAIRY, MALE, LIVE	
BOVINES, PUREBRED BREEDING, DAIRY, FEMALE, LIVE	
BOVINES, PUREBRED BREEDING, MALE, LIVE, EXCEPT DAIRY	
BOVINES, PUREBRED BREEDING, FEMALE, LIVE, EXCEPT DAIRY	
BOVINES, LIVE, NESOI	
SWINE, PUREBRED BREEDING, LIVE	
SWINE, WEIGHING LESS THAN 50 KG EACH, LIVE, EXCEPT PUREBRED BREEDING	
. . .	

4

Search changed to filter out 'FEMALE'

```

SELECT col1
  FROM customs_data
  WHERE
        col1 LIKE '%PUREBRED%'
      AND
        col1 LIKE '%MALE%'
-----+-----+-----+-----+-----+-----+-----+
COL1
-----+-----+-----+-----+-----+-----+-----+

DSNE610I NUMBER OF ROWS DISPLAYED IS 0

```

Adding more search terms

```

SELECT col1
  FROM customs_data
  WHERE
        col1 LIKE '%PUREBRED%'
      AND
        col1 LIKE '%MALE%'
      AND
        col1 NOT LIKE '%FEMALE%'
-----+-----+-----+-----+-----+-----+-----+
CMY_CLS_SCH_B_TE
-----+-----+-----+-----+-----+-----+-----+
BOVINES, PUREBRED BREEDING, DAIRY, MALE, LIVE
BOVINES, PUREBRED BREEDING, MALE, LIVE, EXCEPT DAIRY

DSNE610I NUMBER OF ROWS DISPLAYED IS 2

```

Drawbacks with LIKE

- Searches a string of bytes for a pattern match, therefore:
 - is case sensitive
 - doesn't understand language delimiters: “,” “.” “ ”
 - doesn't do language processing: “have / has / had”, “male” vs. “female”, English vs. Spanish
 - can't do synonym processing: “bovine / cow / cattle / bull”
 - coding can be challenging for multiple search terms

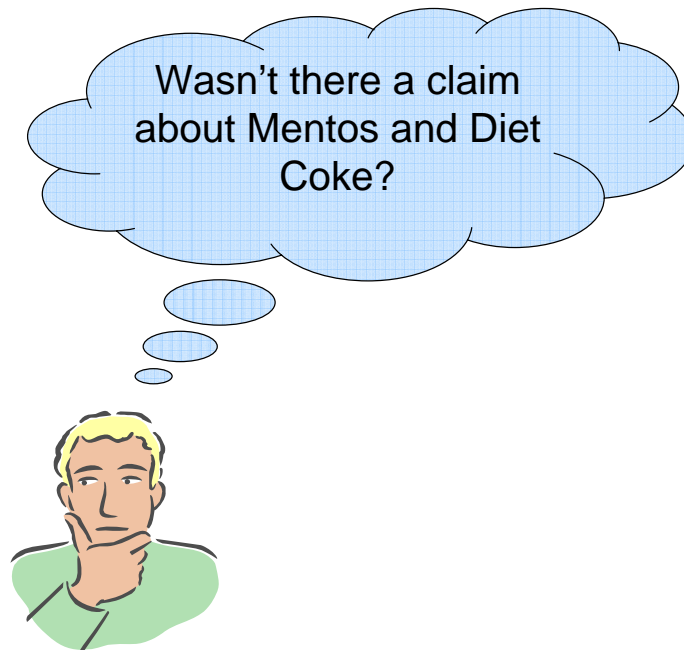
11

Customer search scenarios

- DB2 for z/OS table with catalog item descriptions
 - VARCHAR column, average 256 bytes
 - Online web searching with familiar interface
- DB2 for z/OS table with insurance agent notes
 - CLOB column, average 1K bytes
 - Agent remembers a claim but not who made it!
- DB2 for z/OS table with item names
 - CHAR column, padded to 80 bytes, 400K+ rows
 - Find items with keyword ordered by a customer

12

Example of a claim question



13

Customer demand for text search

- Common Scenario:
 - Text data in DB2 for z/OS: catalog, reports
 - Need online capability for users to search
 - Data extracted from DB2 to enable search
 - Search and retrieve are separate operations
- Customers want to use DB2 SQL API for a single search and retrieve
 - Security, optimization

14

Problems with prior search solutions

- Early attempt was “Text Extender”, with many technical and usability problems
 - Not integrated with DB2 for z/OS – a “wart” on top
 - Difficult for customers to manage, use of file system, etc.
 - Performance limited by external functions
 - Index size scalability issues
 - Lack of data sharing support
 - Used z/OS MIPS with associated MLC woes
- Newer “Net Search Extender” was not ported to z/OS

15

DB2 for z/OS requirements

- All administration from invoking administrative stored procedure
 - No UNIX System Services commands
 - Use standardized DB2 infrastructure and interfaces
 - Admin SPs defined by DB2 installation JCL
 - No new authorizations, ids, password files, etc.
- No HFS file system access from DB2 engine or admin SP
 - All data in DB2 tables
 - Administrative info accessible using SQL

16

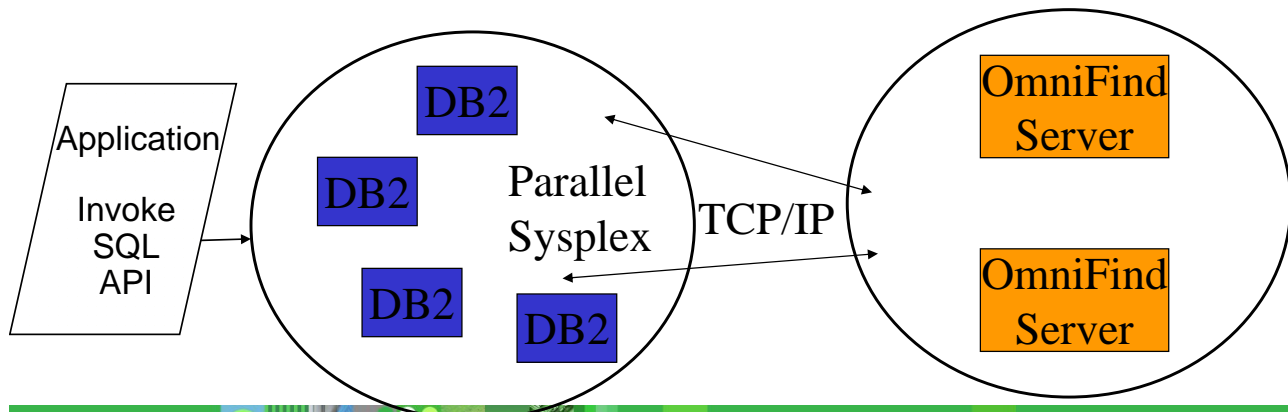
DB2 for z/OS requirements

- Must support data sharing group
 - All data and indexes must be accessible from all members
- Backup of text search server index data must be able to be coordinated with DB2 data
- Must be able to support failover to a different text search server

Architecture

OmniFind Text Search support

- Provides text search for CHAR/VARCHAR/LOB/XML columns
- OmniFind provides a text index server
- Efficient communication interactions with DB2 for z/OS
- OmniFind text indexes are persisted into DB2 tables for backup/recovery purposes



19

What kind of data can I search?

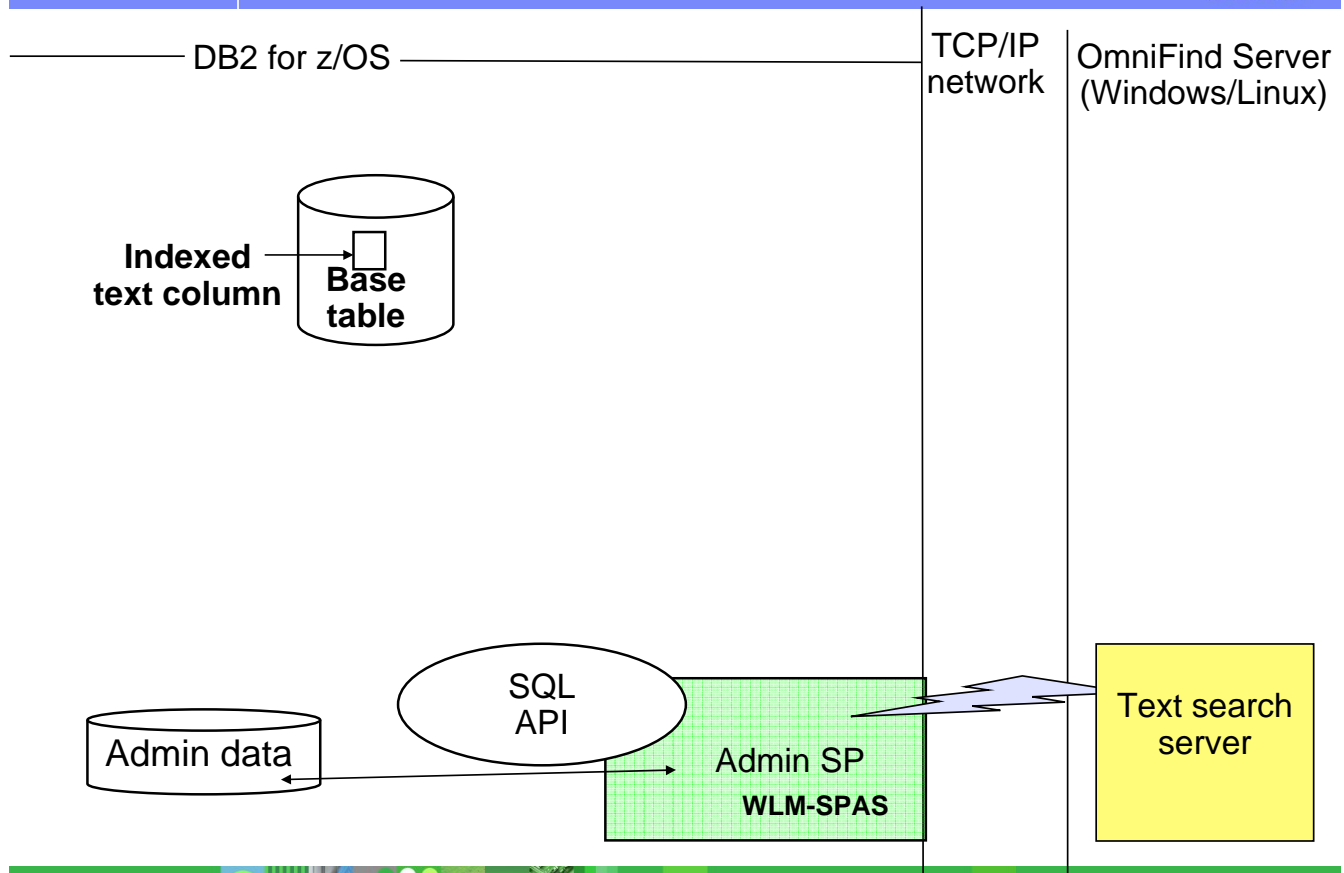
- **Character**
 - CHAR, VARCHAR, CLOB
 - Encoding CCSID known by DB2
 - Language can be specified
- **BLOB, FOR BIT DATA, (VAR)BINARY**
 - rich text: PDF, DOC, PPT, etc
 - CCSID, language can be specified
- **XML**
 - Provide for ad-hoc Xpath expressions

20

Text search server

- IBM OmniFind Text Search Server for DB2 for z/OS
 - Doesn't that just roll off of your tongue?
- Windows or Linux server required
 - Linux: Red Hat or SUSE
 - not z/Linux yet: please **shout** if you require this
 - Windows 2003 server
- Server code shipped with DB2 Accessories Suite for z/OS (5655-R14), at no cost
- Quick install via shell scripts

21

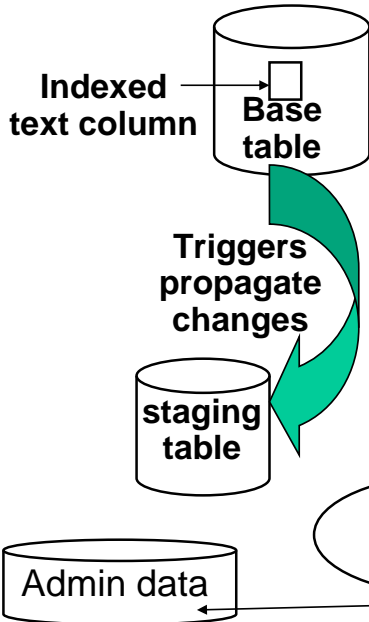


22

DB2 for z/OS

TCP/IP network

OmniFind Server (Windows/Linux)

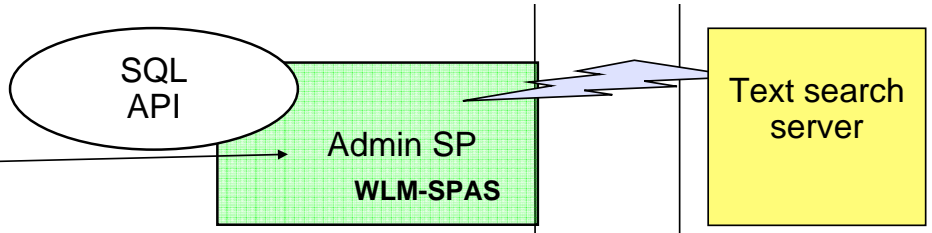


When a text search index is created with the SYSTS_CREATE stored procedure,

A staging table is created and admin data written

A trigger is created on the base table to propagate changes to a staging table

An index collection is created on the text search server (but not populated yet).



DB2 for z/OS

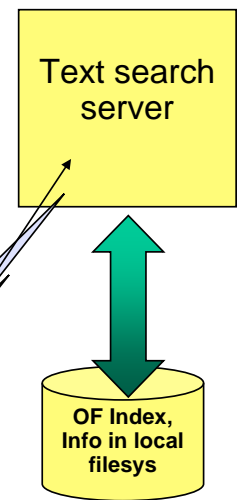
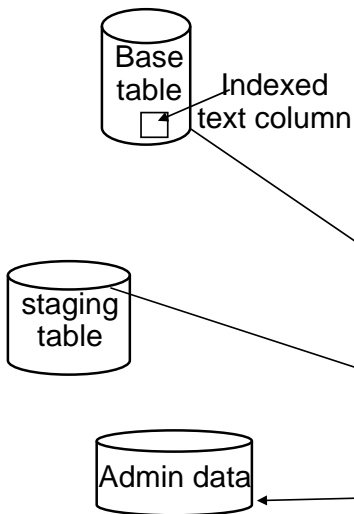
TCP/IP network

OmniFind Server (Windows/Linux)

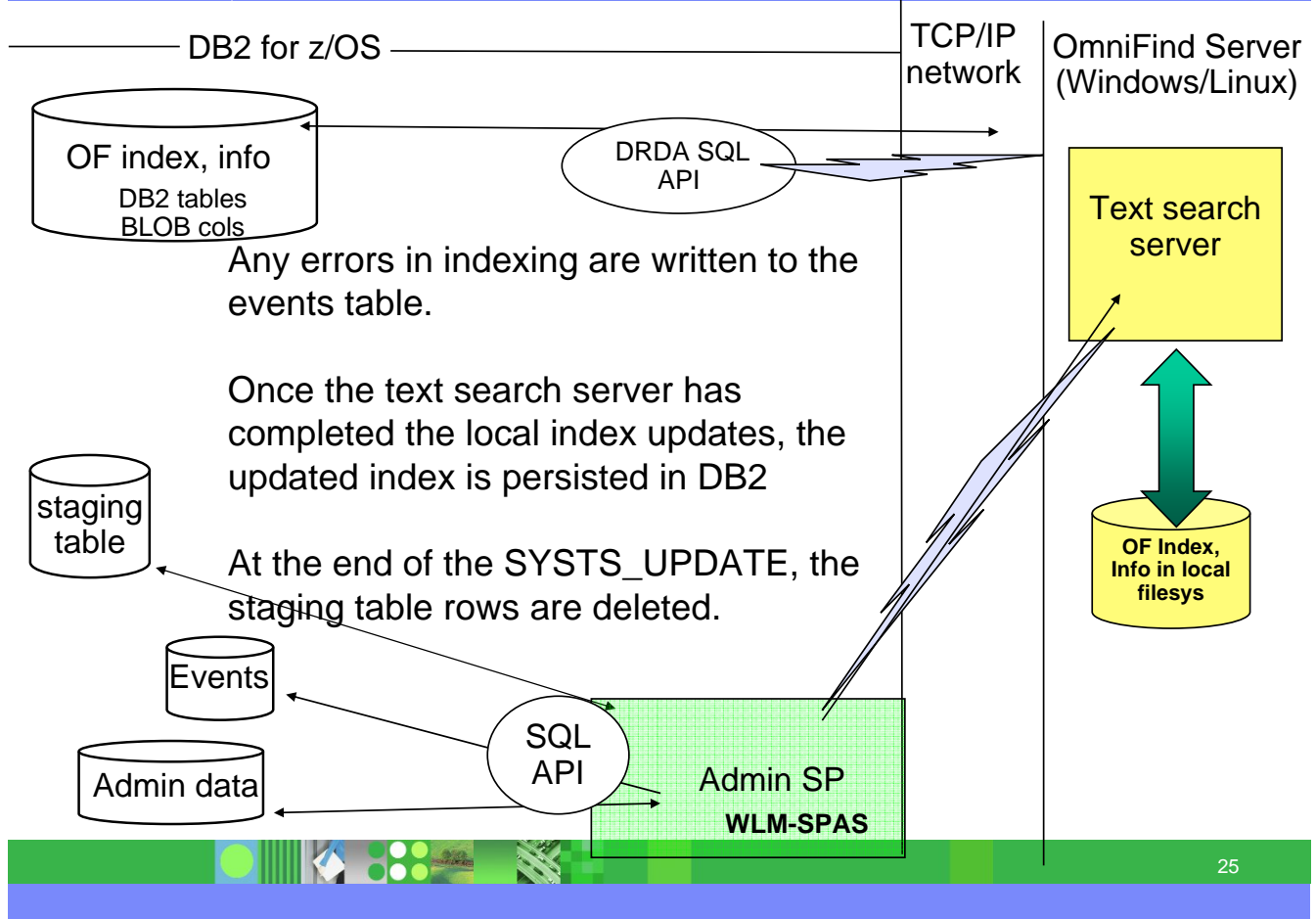
When a text search index is updated using the SYSTS_UPDATE SP,

The staging table is read to find changes, and the changed values are retrieved from the base table and sent to the text search server.

...(con't)



DB2 DS shared data



Searching

Using a text index

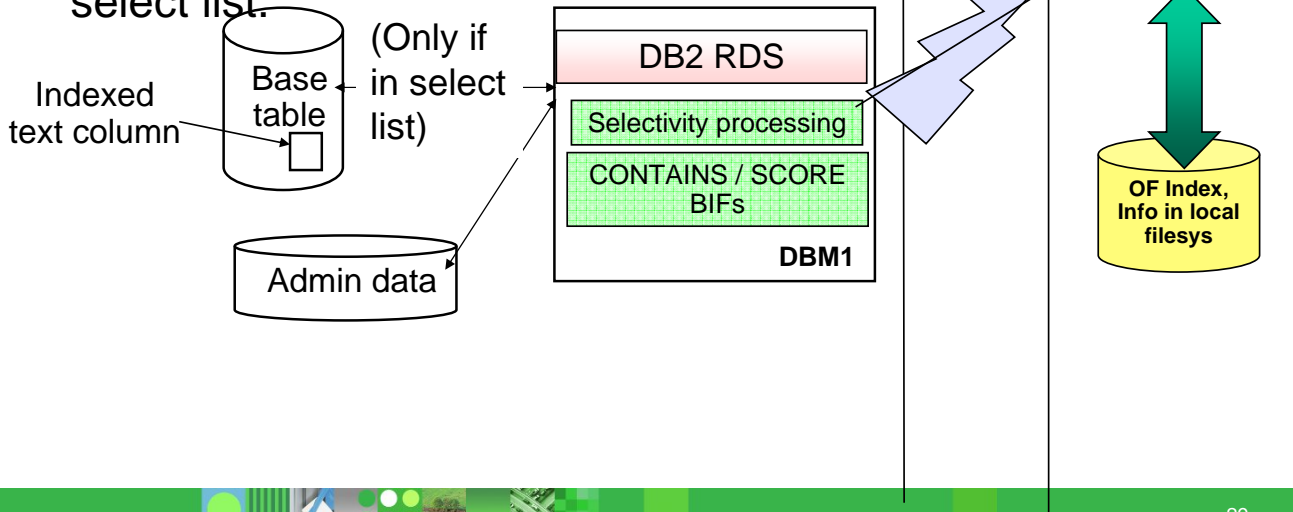
- Built-in functions CONTAINS() and SCORE ()
- Optimizer does selectivity evaluations on built-in functions
 - For example, a more complex WHERE clause may be able to use a DB2 index more efficiently – for example narrowing by state.

```
SELECT CUSTOMER FROM CLAIM_TABLE
WHERE
    CONTAINS( REPORT , 'mentos diet coke ') = 1
    AND STATE = 'CA' ;
```

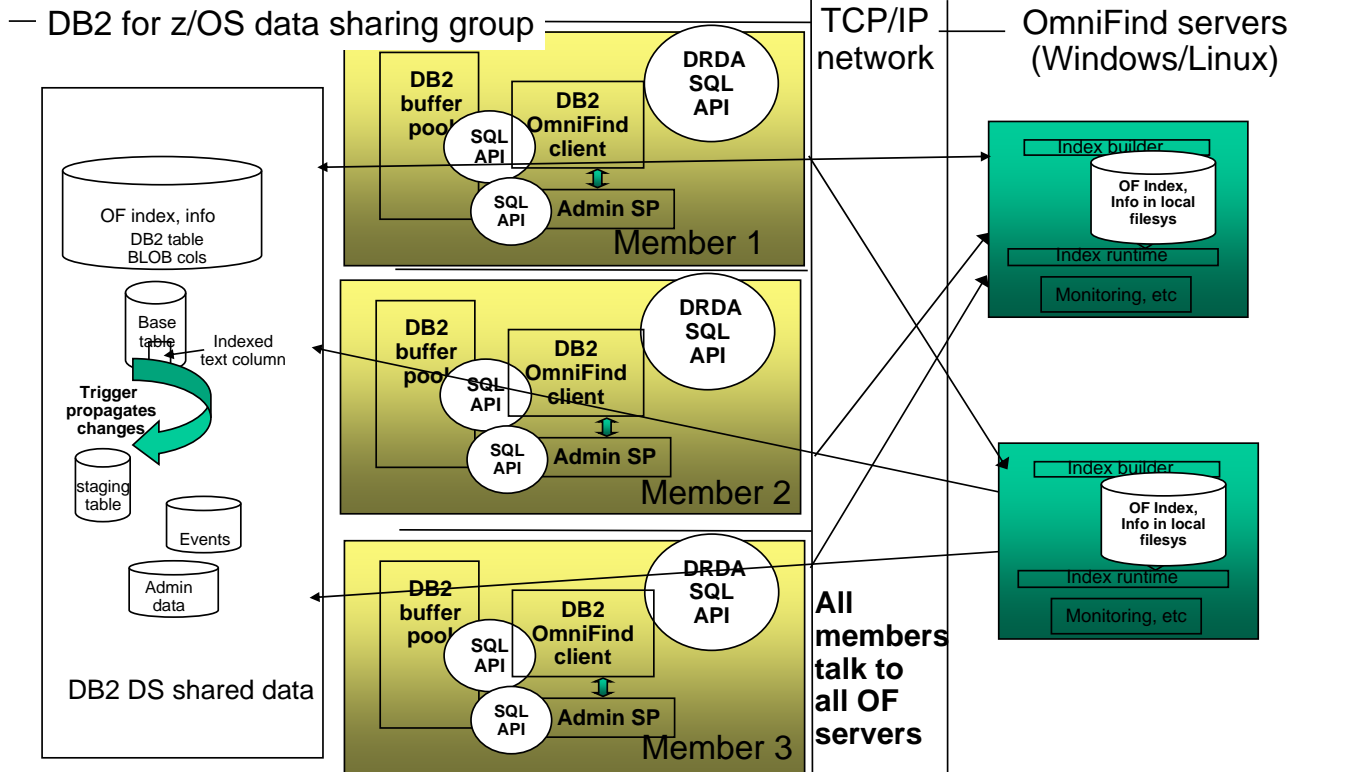
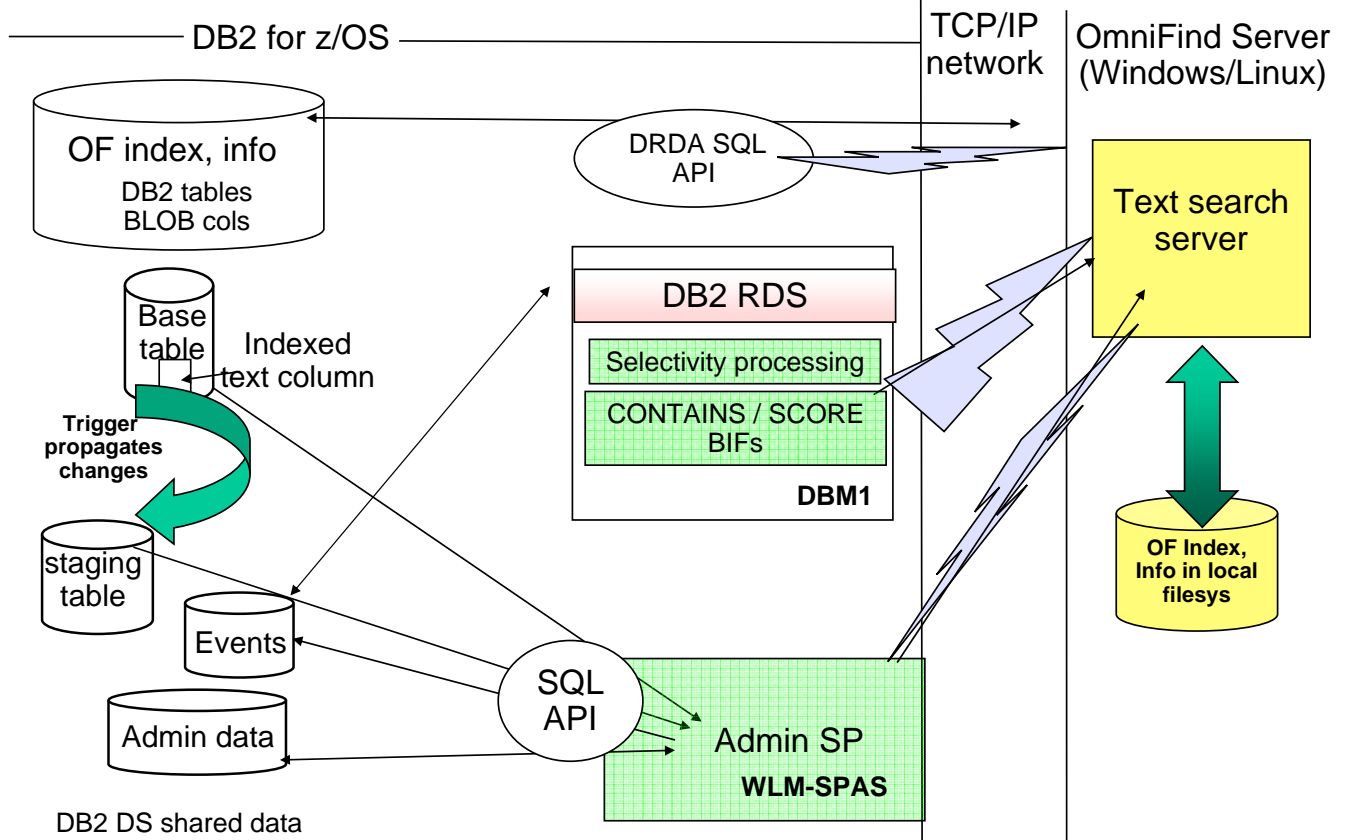
What kinds of searching can I do?

- Familiar operators: +, -, OR, AND, “diet coke”
 - CONTAINS(ITEM_DESC , 'plastic recyclable -bottle')
 - CONTAINS(REPORT , 'flood -plumbing')
- Wildcarding: * , ?
 - CONTAINS(ITEM_DESC , 'poly* roll')
- Grouping: ()
 - CONTAINS(REPORT , 'flood (river OR creek)')

DB2 for z/OS
 For CONTAINS() and SCORE(),
 DB2 contacts the text search
 server for the matching ROWIDs,
 and fetches the base table rows
 only if the indexed column is in the
 select list.



Administration



Key tasks for enabling text search

1. Obtain and install the OmniFind Text Search Server on a Windows or Linux machine
2. Run the installation JCL in DB2 for z/OS to define the stored procedures and create the tables
3. Populate the administration table with the OmniFind server information
 - Location, user/pw, DB2 user/pw
4. Install and configure the type-4 Universal Java Driver on the OmniFind text search server machine
5. Execute the START() stored procedure

33

Key tasks for administering text search

- Choose the text columns to enable indexing
 - CHAR, VARCHAR, CLOB, BLOB, XML
 - Table must have a ROWID column
- Invoke the SYSPROC.SYSTS_CREATE() stored procedure to create the index for each column
 - language may be specified
- Invoke the SYSPROC.SYSTS_UPDATE() stored procedure to index the documents
- Schedule the SYSPROC.SYSTS_UPDATE() stored procedure to run periodically
- Check and clear the events table periodically

34

Text Search administration stored procedures

- Start – DB2 connects to OF server, allows CONTAINS() and SCORE() to be invoked
- Create text index
- Update index
- Alter index, drop index
- Stop OF – CONTAINS() and SCORE() disabled
 - For emergency use only?

Error handling

- Server errors exposed in both SQLCA and operator console messages
- SYSPROC.SYSTS_TAKEOVER stored procedure to try again and move the index handling to another available server
- SYSPROC.SYSTS_RESTORE for planned server outages or recovering DB2 to point in time

Sounds great! When can I have it?

- Available now!
- Order DB2 9 for z/OS and the DB2 Accessories Suite

37

Summary: key points to know about text search

1. Requires OmniFind server distributed through the DB2 Accessories Suite (5655-R14)
2. Number of indexes on one server is limited by memory and disk space
3. Text indexes are not automatically updated by DB2
 - Base table changes may not immediately be reflected
 - Must invoke `SYSPROC.SYSTS_UPDATE`
 - Can be automated through the DB2 Administration Console
4. Text indexes are not dropped when table is dropped

38

Oh, about that query?

```
SELECT CUSTOMER FROM CLAIM_TABLE  
WHERE CONTAINS ( REPORT , "mentos diet  
coke " );
```

Huh?

Why would an accident report contain
both Mentos (candy) and Diet Coke?

39



More at: YouTube -- search on 'mentos "diet coke" '
• the video from "Mythbusters" is good!

40

If you went to IOD . . .

- Take a look at Kevin Campbell's (Univar USA) presentation about his experiences with DB2 for z/OS Text Search:

- Session 1075 – “User Experiences with OmniFind Text Search Server at Univar USA”

“Univar USA was an early adopter of IBM OmniFind™ Text Search Server. This presentation details the business problem that OmniFind solved and provides detailed notes on usage and performance.”

41

Disclaimer and Trademarks

Information contained in this material has not been submitted to any formal IBM review and is distributed on "as is" basis without any warranty either expressed or implied. Measurements data have been obtained in laboratory environment. **Information in this presentation about IBM's future plans reflect current thinking and is subject to change at IBM's business discretion. You should not rely on such information to make business plans. The use of this information is a customer responsibility.**

IBM MAY HAVE PATENTS OR PENDING PATENT APPLICATIONS COVERING SUBJECT MATTER IN THIS DOCUMENT. THE FURNISHING OF THIS DOCUMENT DOES NOT IMPLY GIVING LICENSE TO THESE PATENTS.

TRADEMARKS: THE FOLLOWING TERMS ARE TRADEMARKS OR ® REGISTERED TRADEMARKS OF THE IBM CORPORATION IN THE UNITED STATES AND/OR OTHER COUNTRIES: AIX, AS/400, DATABASE 2, DB2, e-business logo, Enterprise Storage Server, ESCON, FICON, OS/390, OS/400, ES/9000, MVS/ESA, Netfinity, RISC, RISC SYSTEM/6000, iSeries, pSeries, xSeries, SYSTEM/390, IBM, Lotus, NOTES, WebSphere, z/Architecture, z/OS, zSeries, OmniFind



The FOLLOWING TERMS ARE TRADEMARKS OR REGISTERED TRADEMARKS OF THE MICROSOFT CORPORATION IN THE UNITED STATES AND/OR OTHER COUNTRIES: MICROSOFT, WINDOWS, WINDOWS NT, ODBC, WINDOWS 95

For additional information see ibm.com/legal/copytrade.phtml

1

42

Searching Unstructured Text in DB2 for z/OS: When LIKE Isn't Good Enough

Mark Wilson

IBM

Email: mwilson2@us.ibm.com

Thank You!!!