Source Code Search

and other applications of in-memory technology in software development

Oleksandr Panchenko

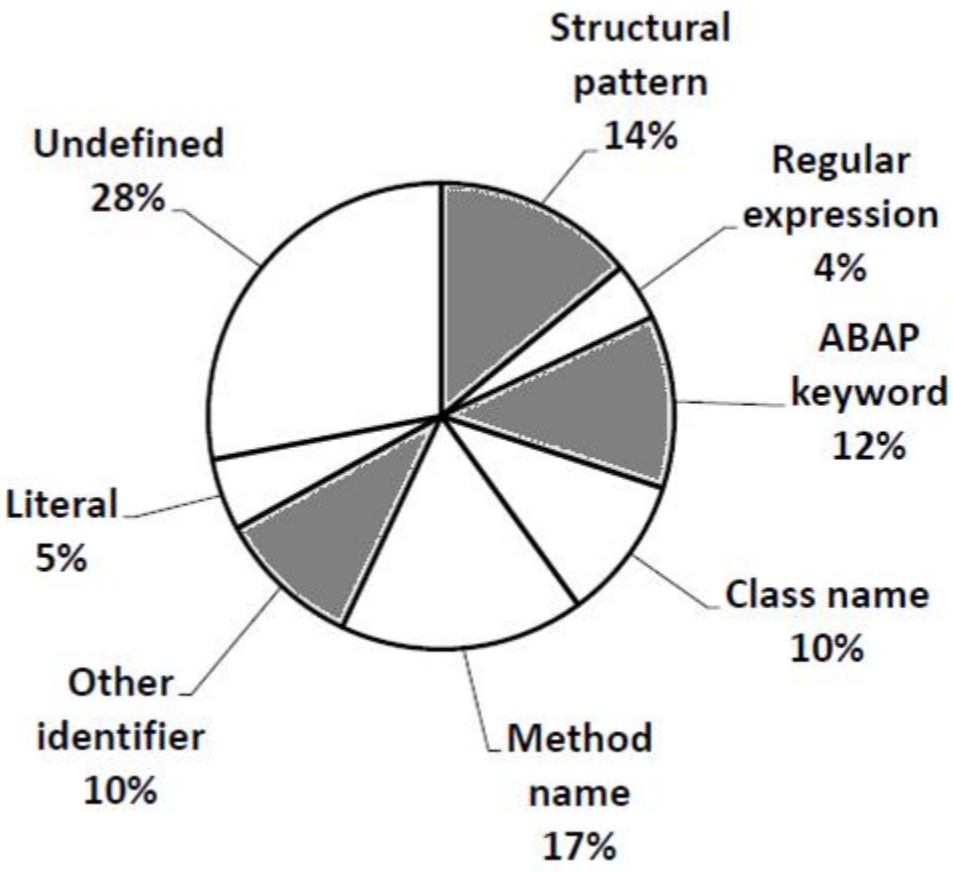
29.06.2011

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

There is a huge amount of source code:

- at SAP:
  - Business ByDesign and NetWeaver – 72 Mio ABAP statements, 124k development objects (classes, reports, function modules), 1,6 Mio includes
  - NetWeaver Basis – 12k database tables, 800k SQL statements
  - Business Suite – 400 Mio lines of code
- at Microsoft:
  - Microsoft Vista – 120 Mio lines of code
- at Google:
  - 20+ Mio lines of code and 900k projects – released for public access
- on the Web:
  - SourceForge.net – 295k open source projects
  - Google Project Hosting – 250k open source projects

Categories of Developers' Queries



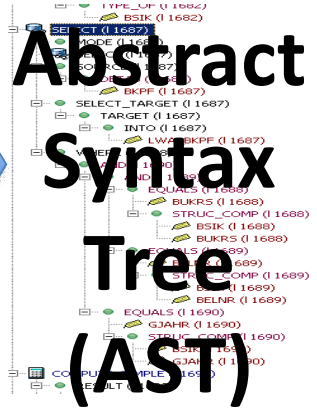
Example search targets:

1. compute statements that assign a value to a global variable
2. updating access to a database table
3. violation of some development guidelines
4. is a modification of (3) with an identifier added
5. all function invocations which use, e.g., `param_name`, as a formal parameter name
6. all occurrences of a system/global variable, e.g., `uname`, in an ASSERT statement
7. data declarations of a certain type
8. authorization checks

```
SELECT SINGLE * FROM bkpf INTO lwa_bkpf
    WHERE bukrs = bsik-bukrs
    AND   belnr = bsik-belnr
    AND   gjahr = bsik-gjahr.

ktab-xblnr_alt = lwa_bkpf-xblnr_alt.

*      Invoices and credit memos
  IF bsik-umsks IS INITIAL AND
     bsik-zumsk IS INITIAL AND
* Note 512749 - BoI
*      bsik-reb
* Note 512749 - BoI
* Note 512749 - BoI
     bsik-xsahl IS INITIAL AND
     bsik-bschl IN buschl.
* Note 512749 - EoI
    APPEND ktab TO ewt_invoice_tab.
  ENDIF.

*      Downpay
  IF bsik-umsks
     bsik-zumsk IS INITIAL AND
     bsik-rebzj IS INITIAL AND
     bsik-rebzg IS INITIAL AND
     bsik-rebzz IS INITIAL AND
     bsik-rebzt IS INITIAL.
    APPEND ktab TO ewt_downpay_tab.
  ENDIF.

*      Downpayment clearings (part pointing
```

Source code

1

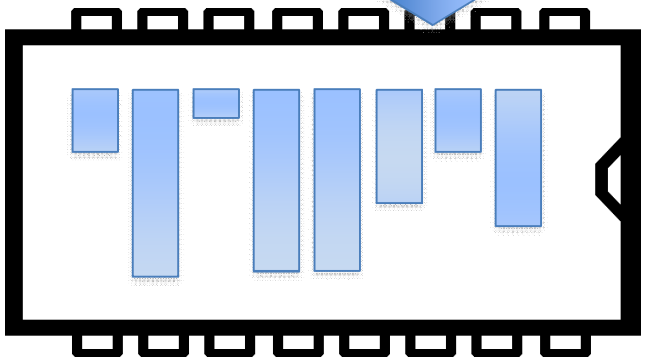


Abstract Syntax Tree (AST)

2

	Attribute	# of Rows	Cardinality
PK	DEV_OBJECT_NAME	458M	124k
	VERTEX_VALUE	458M	5,7M
	VERTEX_CATEGORY	458M	6
PK	PRE_ORDER	458M	5,3M
	POST_ORDER	458M	5,3M
	INCLUDE_NAME	458M	1,6M
	SOURCE_POSITION	458M	115k
FK	PARENT_PRE_ORDER	458M	3,4M

3



4

```
CALL FUNCTION …
EXPORTING simplified = …
```

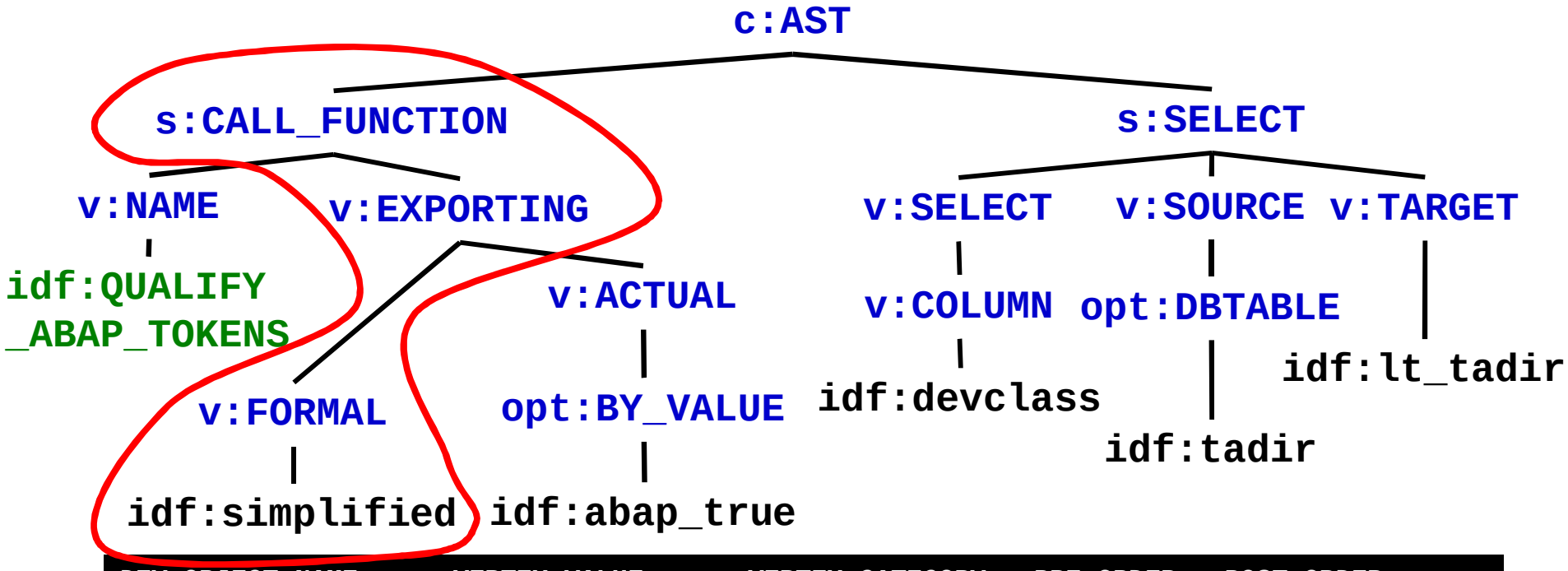
5

```
//s:CALL_FUNCTION
[./v:EXPORTING/v:FORMAL/idf:simplified]
```

6

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

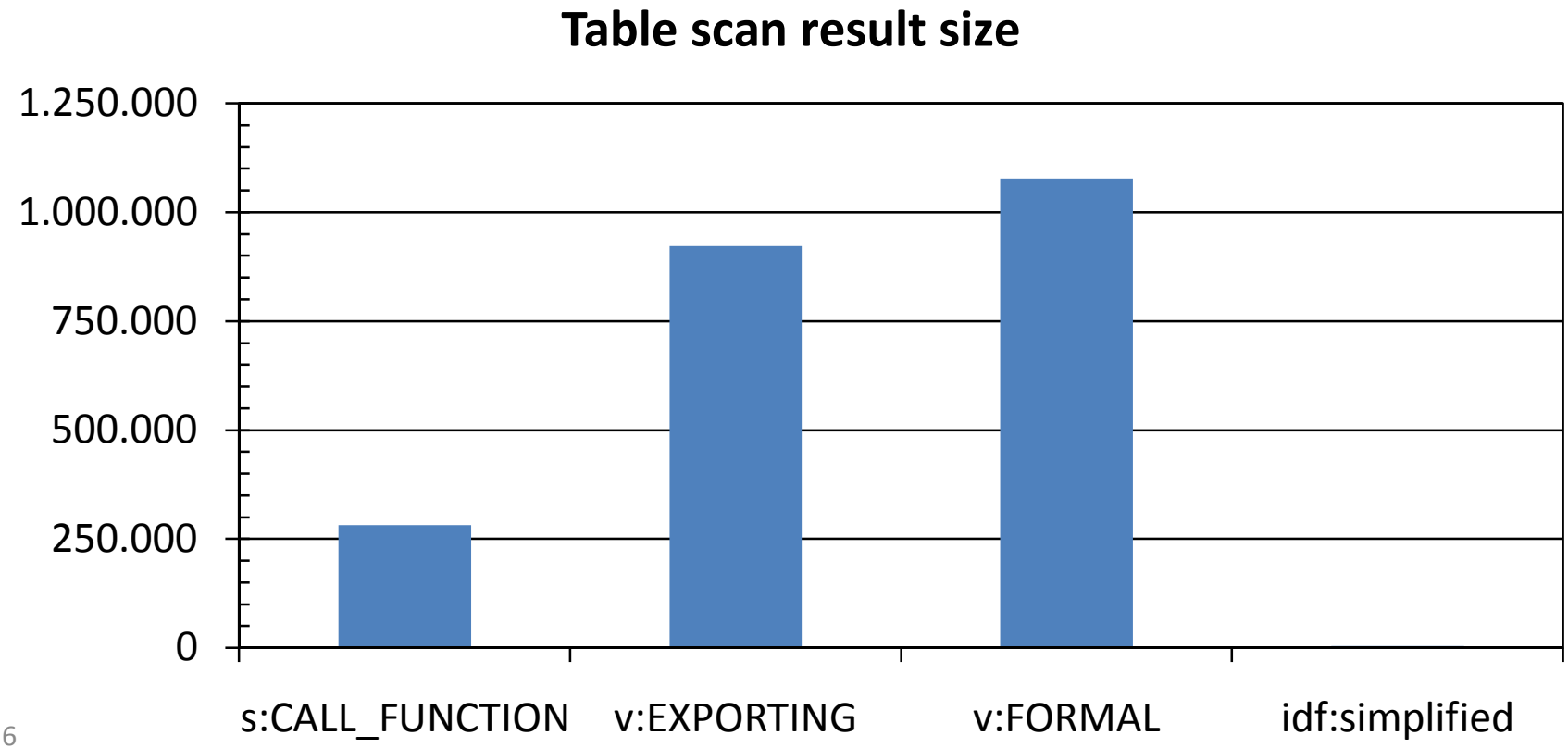
```
CALL FUNCTION 'QUALIFY_ABAP_TOKENS'
    EXPORTING simplified = abap_true.
SELECT devclass FROM tadir INTO lt_tadir.
```



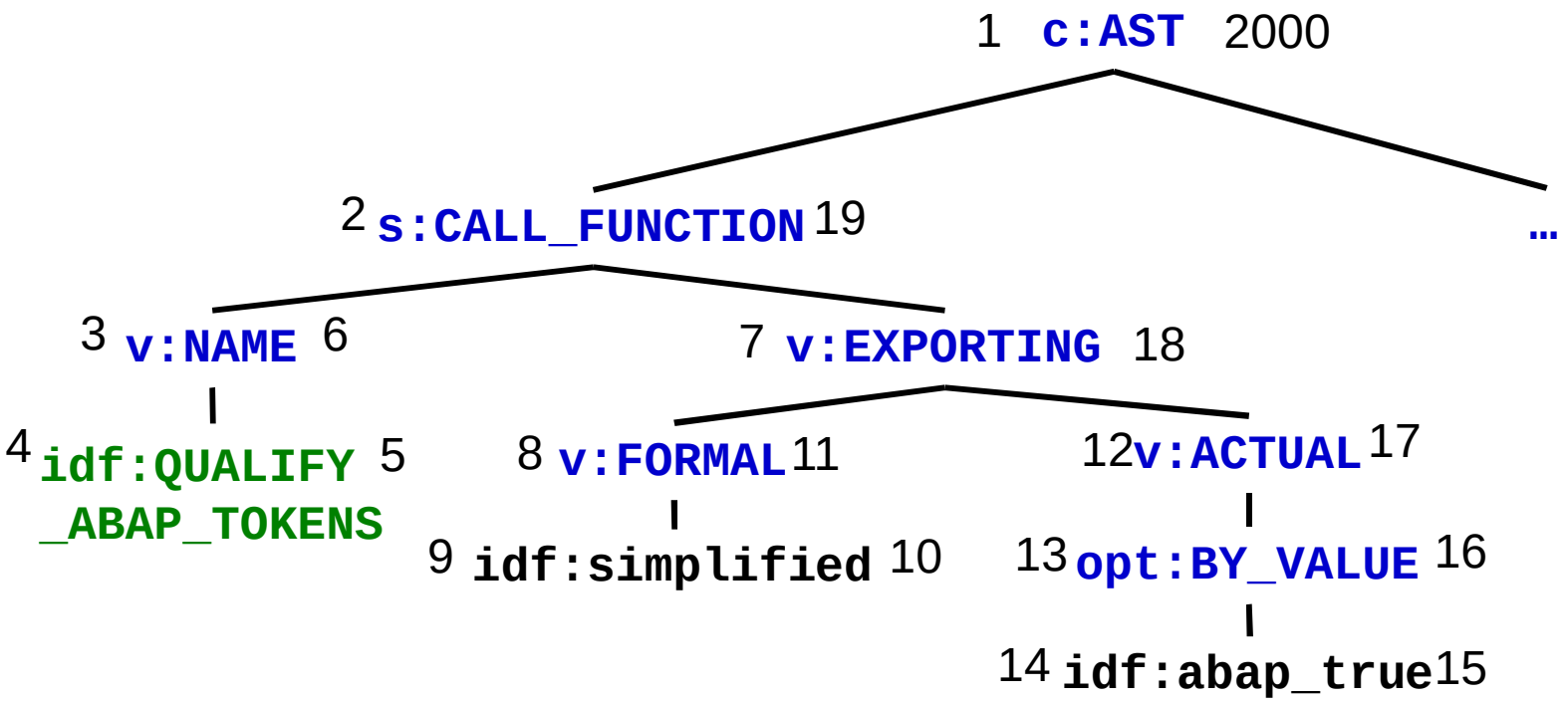
DEV_OBJECT_NAME	VERTEX_VALUE	VERTEX_CATEGORY	PRE_ORDER	POST_ORDER	…
CL_FREE_SEARCH	AST	compound	1	38	…
CL_FREE_SEARCH	CALL_FUNCTION	statement	2	19	…
CL_FREE_SEARCH	NAME	clause	3	6	…
CL_FREE_SEARCH	QUALIFY_ABAP_TOKENS	literal	4	5	…
CL_FREE_SEARCH	EXPORTING	clause	7	18	…
CL_FREE_SEARCH	FORMAL	clause	8	11	…
CL_FREE_SEARCH	simplified	identifier	9	10	…
…	…	…	…	…	…

Query Execution

```
//s:CALL_FUNCTION[./v:EXPORTING/v:FORMAL/idf:simplified]
```

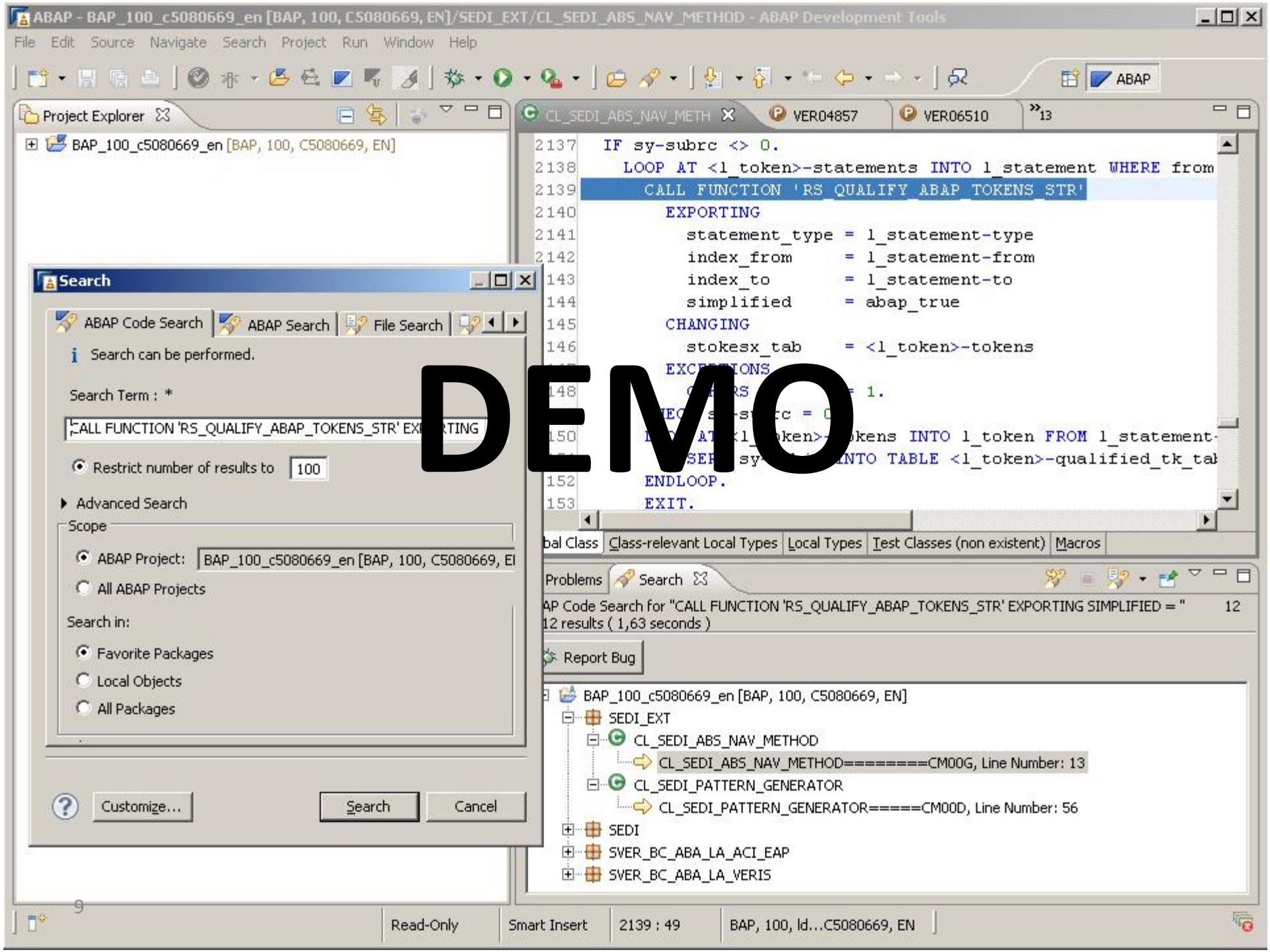


Pre- Post-Order Numbering



Why In-Memory?

- **Since the number of AST vertex types is limited, the database can be well compressed**
  - **3,2GB of original source code in text files**
  - **80,0GB of AST data because of parsing**
  - **8,5GB compressed AST data**
- **Special indexing allows faster query evaluation**
  - **Pre-post-order numbering**
  - **Set processing instead of traversal of ASTs**
  - **Parallel data processing**

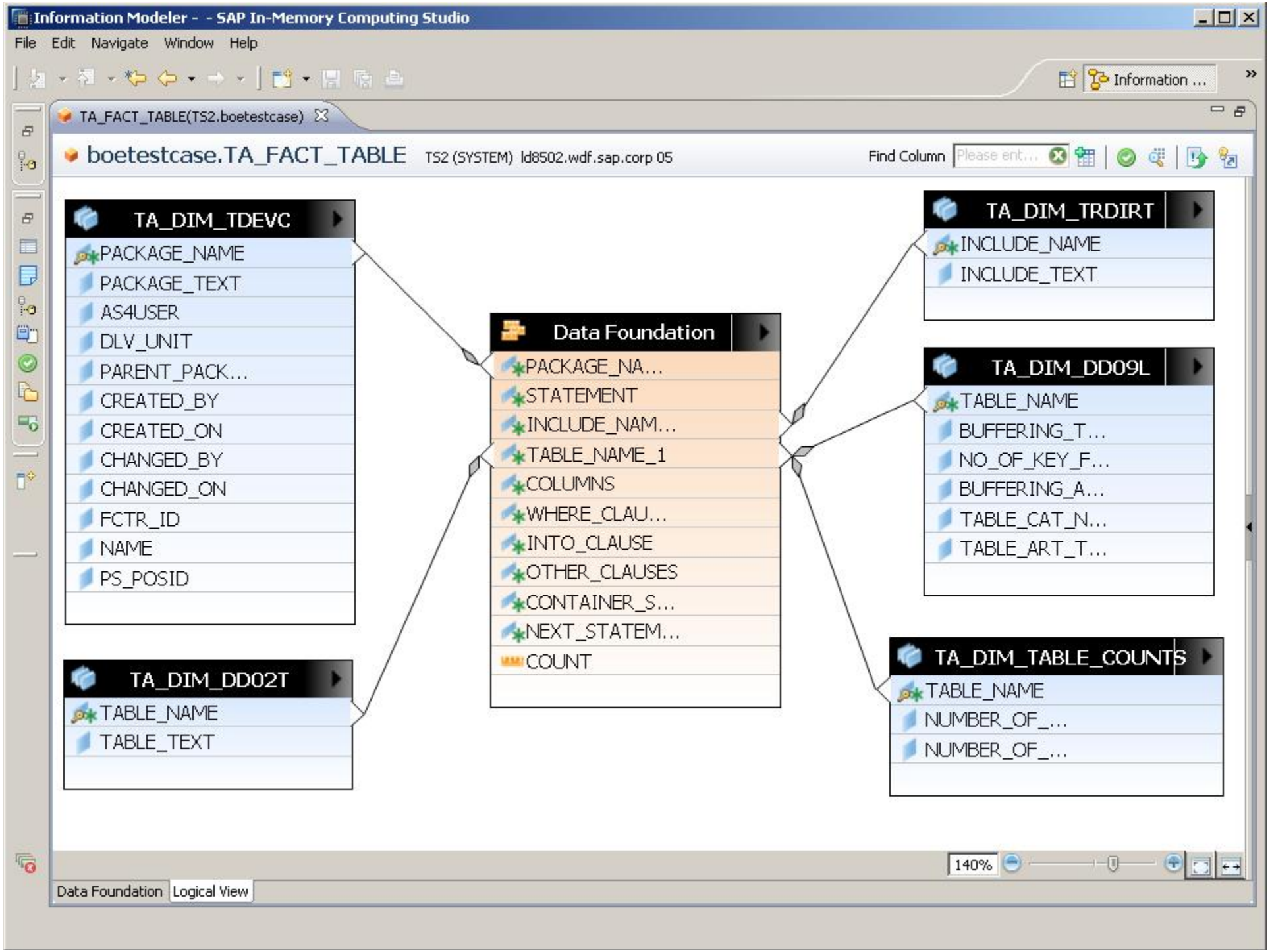

ABAP - BAP_100_c5080669_en [BAP, 100, C5080669, EN]/SEDI_EXT/CL_SEDI_ABS_NAV_METHOD - ABAP Development Tools
File Edit Source Navigate Search Project Run Window Help
ABAP
Project Explorer
BAP_100_c5080669_en [BAP, 100, C5080669, EN]
CL_SEDI_ABS_NAV_METH
VER04857
VER06510
2137 IF sy-subrc <> 0.
2138 LOOP AT <l_token>-statements INTO l_statement WHERE from
2139 CALL FUNCTION 'RS_QUALIFY_ABAP_TOKENS_STR'
2140 EXPORTING
2141 statement_type = l_statement-type
2142 index_from = l_statement-from
143 index_to = l_statement-to
144 simplified = abap_true
145 CHANGING
146 stokesx_tab = <l_token>-tokens
148 = 1.
150 INTO l_token FROM l_statement-
152 ENDLOOP.
153 EXIT.
Class-relevant Local Types
Local Types
Test Classes (non existent)
Macros
Search
ABAP Code Search
ABAP Search
File Search
Search can be performed.
Search Term : *
CALL FUNCTION 'RS_QUALIFY_ABAP_TOKENS_STR' EXP TING
Restrict number of results to 100
Advanced Search
Scope
ABAP Project: BAP_100_c5080669_en [BAP, 100, C5080669, EI
All ABAP Projects
Search in:
Favorite Packages
Local Objects
All Packages
Customize...
Search
Cancel
DEMO
Problems
Search
AP Code Search for "CALL FUNCTION 'RS_QUALIFY_ABAP_TOKENS_STR' EXPORTING SIMPLIFIED = "
12 results (1,63 seconds)
Report Bug
BAP_100_c5080669_en [BAP, 100, C5080669, EN]
SEDI_EXT
CL_SEDI_ABS_NAV_METHOD
CL_SEDI_ABS_NAV_METHOD========CM00G, Line Number: 13
CL_SEDI_PATTERN_GENERATOR
CL_SEDI_PATTERN_GENERATOR=====CM00D, Line Number: 56
SEDI
SVER_BC_ABA_LA_ACI_EAP
SVER_BC_ABA_LA_VERIS
Read-Only
Smart Insert
2139 : 49
BAP, 100, ld...C5080669, EN

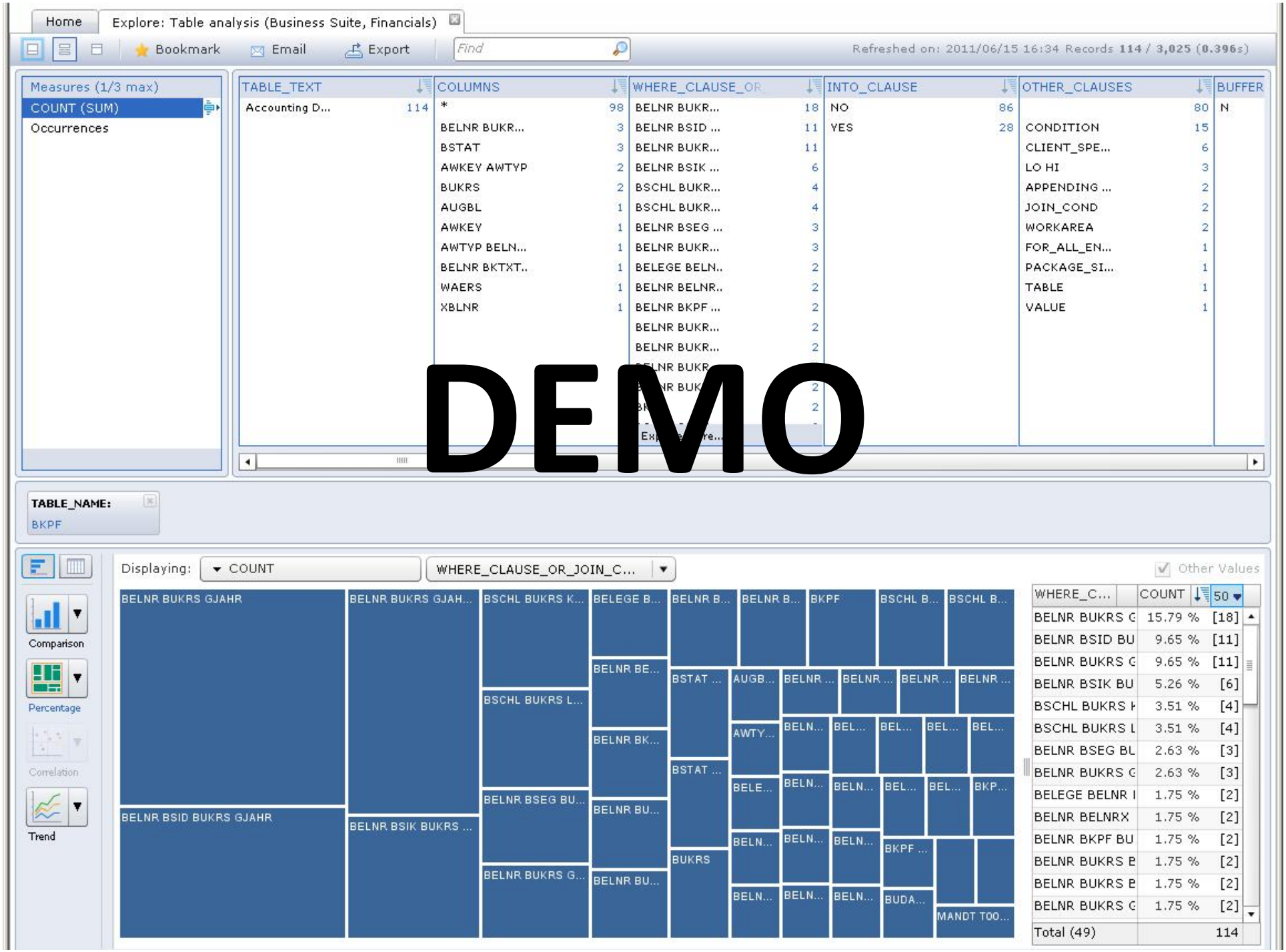
Other Applications

- **Database table usage analysis**
- **Clone detection (visualisation)**
- **Trace analysis**

Database Table Usage Analysis

- **Goal: reengineering of the database layer**
- **For each table identify its usage patterns in source code (footprint):**
  - **In which packages/development objects is used**
  - **Which columns are used (projected, selected, joined, etc.)**
  - **What are technical settings of the table (buffering, sizing, total number of columns, etc.)**
  - **In which context is used (statements around the SQL statement)**
- **Find tables which follow a certain usage pattern, for example, large or wide tables, select statements in a loop, incorrect sizing, etc.**


Information Modeler - - SAP In-Memory Computing Studio
File Edit Navigate Window Help
Information ...
TA_FACT_TABLE(TS2.boetestcase)
boetestcase.TA_FACT_TABLE
TS2 (SYSTEM) ld8502.wdf.sap.corp 05
Find Column
Please ent...
TA_DIM_TDEVC
PACKAGE_NAME
PACKAGE_TEXT
AS4USER
DLV_UNIT
PARENT_PACK...
CREATED_BY
CREATED_ON
CHANGED_BY
CHANGED_ON
FCTR_ID
NAME
PS_POSID
TA_DIM_DD02T
TABLE_NAME
TABLE_TEXT
Data Foundation
PACKAGE_NA...
STATEMENT
INCLUDE_NAM...
TABLE_NAME_1
COLUMNS
WHERE_CLAU...
INTO_CLAUSE
OTHER_CLAUSES
CONTAINER_S...
NEXT_STATEM...
COUNT
TA_DIM_TRDIRT
INCLUDE_NAME
INCLUDE_TEXT
TA_DIM_DD09L
TABLE_NAME
BUFFERING_T...
NO_OF_KEY_F...
BUFFERING_A...
TABLE_CAT_N...
TABLE_ART_T...
TA_DIM_TABLE_COUNTS
TABLE_NAME
NUMBER_OF_...
NUMBER_OF_...
140%
Data Foundation
Logical View


Home
Explore: Table analysis (Business Suite, Financials)
Bookmark
Email
Export
Find
Refreshed on: 2011/06/15 16:34 Records 114 / 3,025 (0.396s)
Measures (1/3 max)
COUNT (SUM)
Occurrences
TABLE_TEXT
COLUMNS
WHERE_CLAUSE_OR
INTO_CLAUSE
OTHER_CLAUSES
BUFFER
DEMO
TABLE_NAME:
BKPF
Displaying:
COUNT
WHERE_CLAUSE_OR_JOIN_C...
Other Values
Comparison
Percentage
Correlation
Trend
BELNR BUKRS GJAHR
BELNR BSID BUKRS GJAHR
BELNR BUKRS GJAH...
BELNR BSIK BUKRS ...
BSCHL BUKRS K...
BSCHL BUKRS L...
BELNR BSEG BU...
BELNR BUKRS G...
BELEGE B...
BELNR BE...
BELNR BK...
BELNR BU...
BELNR BU...
BELNR B...
BELNR B...
BKPF
BSCHL B...
BSCHL B...
BSTAT ...
BSTAT ...
BUKRS
AUGB...
AWTY...
BELE...
BELN...
BELN...
BELN...
BELN...
BELNR ...
BELNR ...
BELNR ...
BELNR ...
BEL...
BEL...
BEL...
BEL...
BEL...
BELN...
BELN...
BELN...
BELN...
BELN...
BELN...
BELN...
BELN...
BKP...
BKPF ...
BUDA...
MANDT TOO...
WHERE_C...
COUNT
50
BELNR BUKRS C 15.79 % [18]
BELNR BSID BU 9.65 % [11]
BELNR BUKRS C 9.65 % [11]
BELNR BSIK BU 5.26 % [6]
BSCHL BUKRS K 3.51 % [4]
BSCHL BUKRS L 3.51 % [4]
BELNR BSEG BL 2.63 % [3]
BELNR BUKRS C 2.63 % [3]
BELEGE BELNR I 1.75 % [2]
BELNR BELNRX 1.75 % [2]
BELNR BKPF BU 1.75 % [2]
BELNR BUKRS B 1.75 % [2]
BELNR BUKRS B 1.75 % [2]
BELNR BUKRS C 1.75 % [2]
Total (49) 114