

FNTSK SKNNK

Fonetisk sökning

Sven Heidorn

Centrala Studiestödsnämnden

sven.heidorn@csn.se



Swedish DB2 User Group 05/01/19

Fonetisk sökning

- Bakgrund
 - CSN gör s.k. SPAR-sökningar på namn hos InfoData för ca. 100 000 kr/månad
 - CSN besitter till 99% själv uppgifterna som eftersöks i SPAR
 - Dom höga kostnaderna, det faktum att CSN redan har de uppgifter man söker samt budgetnerskärningar, leder till utveckling av en sökfunktion mot de egna registren

Fonetisk sökning

- Första problemet.....
 - Som så ofta med namnfält och z/Os används SWEDSORT för att få rätt sortering av ÅÄÖ
 - Namnsökning (fritext) bygger med största sannolikhet på LIKE
 - LIKE är inte indexerbart i kombination med FIELD Procedures, dvs. SWEDSORT
 - DBA´erna stoppar utvecklad lösning

Fonetisk sökning

- Andra problemet
 - Karlsson
 - Karlson
 - Karlzon
 - Carlsson
 - Carlson
 - Carlzon
 - + alla andra namn med samma mönster

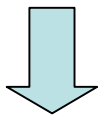
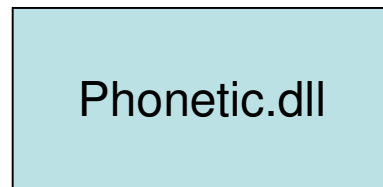
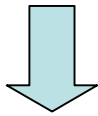
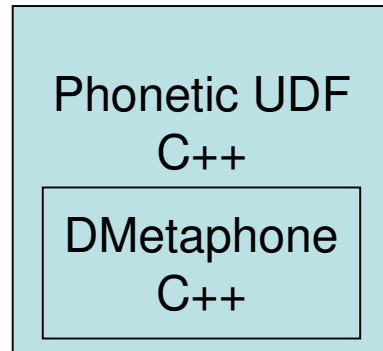
Fonetisk sökning

- Soundex
 - Finns som standard i bl.a. DB2 och MSSQL
 - Ursprungligen utvecklad i slutet av 1800-talet
 - Används i huvudsak för sökning av efternamn, t.ex. vid flygbokningar och släktforskning
 - Fungerar i princip endast med engelska namn

Fonetisk sökning

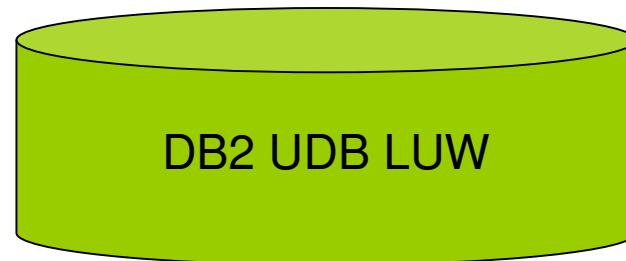
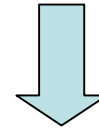
- Metaphone
 - Vidareutveckling av soundex
 - Skapades 1990 av Lawrence Philips
 - Skapades för att kompensera för brister i soundex
 - Vidareutvecklades år 2000 ytterligare till Double Metaphone
 - Används bl.a. i Aspell som är en stavningskontroll

Fonetisk sökning



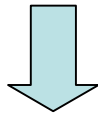
<install path>\SQLLIB\FUNCTION

```
CREATE FUNCTION CSNFUN.PHONETIC (VARCHAR(128))  
  RETURNS VARCHAR(128)  
  EXTERNAL NAME 'Phonetic!Phonetic'  
  LANGUAGE C  
  PARAMETER STYLE SQL  
  DETERMINISTIC  
  FENCED  
  RETURNS NULL ON NULL INPUT  
  NO SQL  
  NO EXTERNAL ACTION  
  NO SCRATCHPAD  
  ALLOW PARALLEL;
```



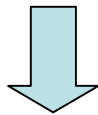
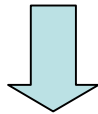
Fonetisk sökning

	Heidorn, Sven	
--	---------------	--

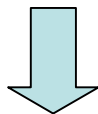


EXPORT to file.ixf of ixf messages msgfile.txt

```
select <other data>, name, <other data>, phonetic(name) from your.userTable
```



LOAD from file.ixf of ixf messages msgfile.txt replace into your.userTable



	Heidorn, Sven		HTRNSFN
--	---------------	--	---------

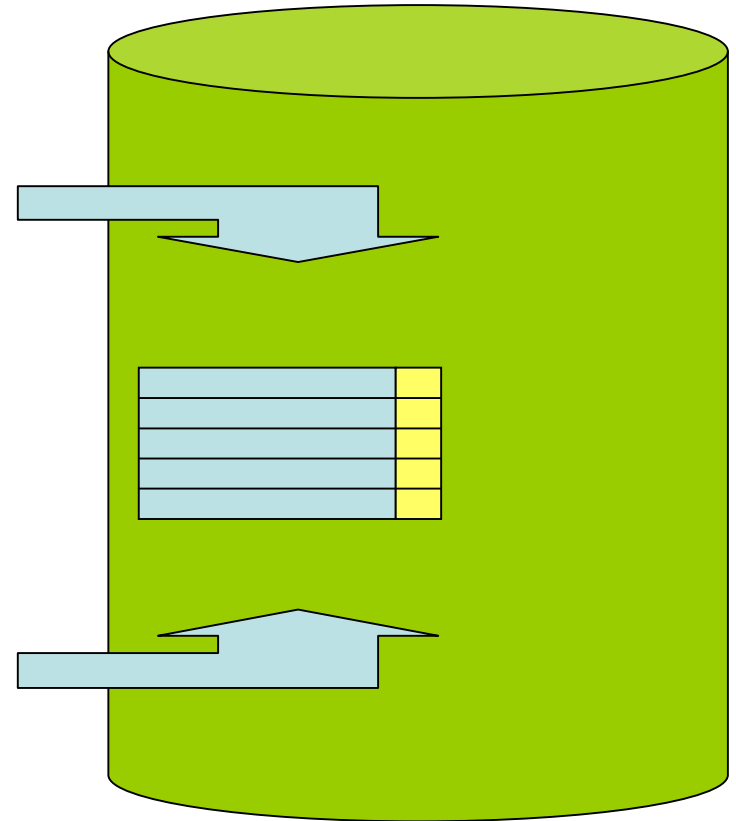
Fonetisk sökning

INSERT

```
CREATE TRIGGER INSTPHON NO CASCADE  
BEFORE INSERT ON your.usrTable  
REFERENCING NEW AS NEW  
FOR EACH ROW MODE DB2SQL  
BEGIN ATOMIC  
  set new.name_phonetic = phonetic(new.name) ;  
END#
```

UPDATE

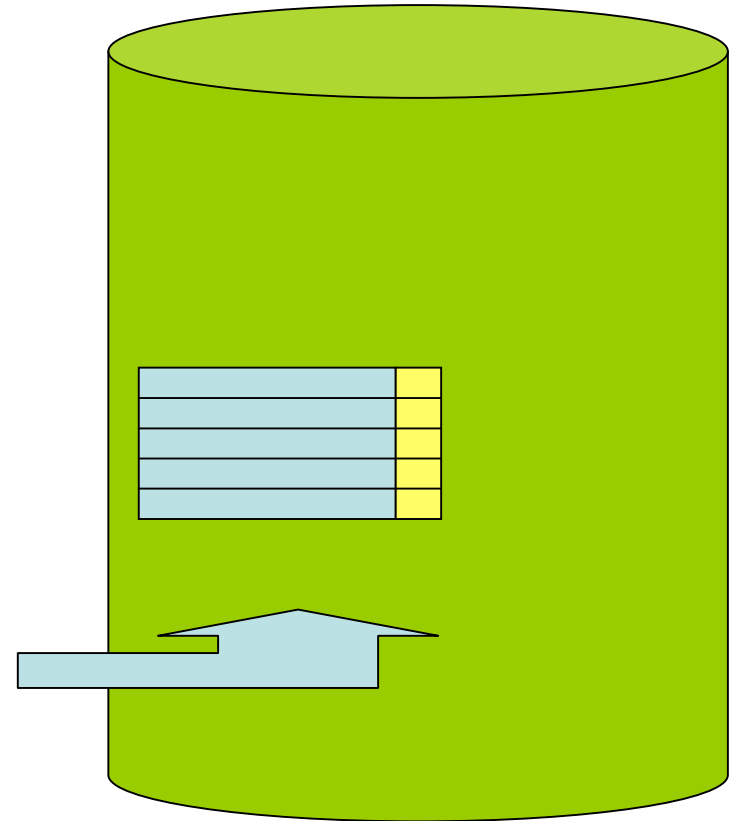
```
CREATE TRIGGER CSNDBA.UPDTPHON NO CASCADE  
BEFORE UPDATE OF NAMN ON your.userTable  
REFERENCING NEW AS NEW  
FOR EACH ROW MODE DB2SQL  
BEGIN ATOMIC  
  set new.name_phonetic = phonetic(new.name) ;  
END#
```



Fonetisk sökning

Applikations Program

```
EXEC SQL  
  SELECT PHONETIC(:hv-name)  
  INTO :hv-name-phonetic:ind-name-phonetic  
  FROM SYSIBM.SYSDUMMY1  
  WITH UR  
  END-EXEC.  
  
EXEC SQL  
  INSERT INTO your.userTable  
  VALUES (<other data>,  
          :hv-name,  
          <other data>,  
          :hv-name-phonetic)  
  END-EXEC
```



Fonetisk sökning

- Double Metaphone kodar bokstäver i strängar
- Övriga tecken ignoreras

```
select rtrim(phonetic('Anka')) || '%' ||  
       rtrim(phonetic('Kalle')) || '%' as soek_straeng  
from sysibm.sysdummy1
```

SOEK_STRAENG

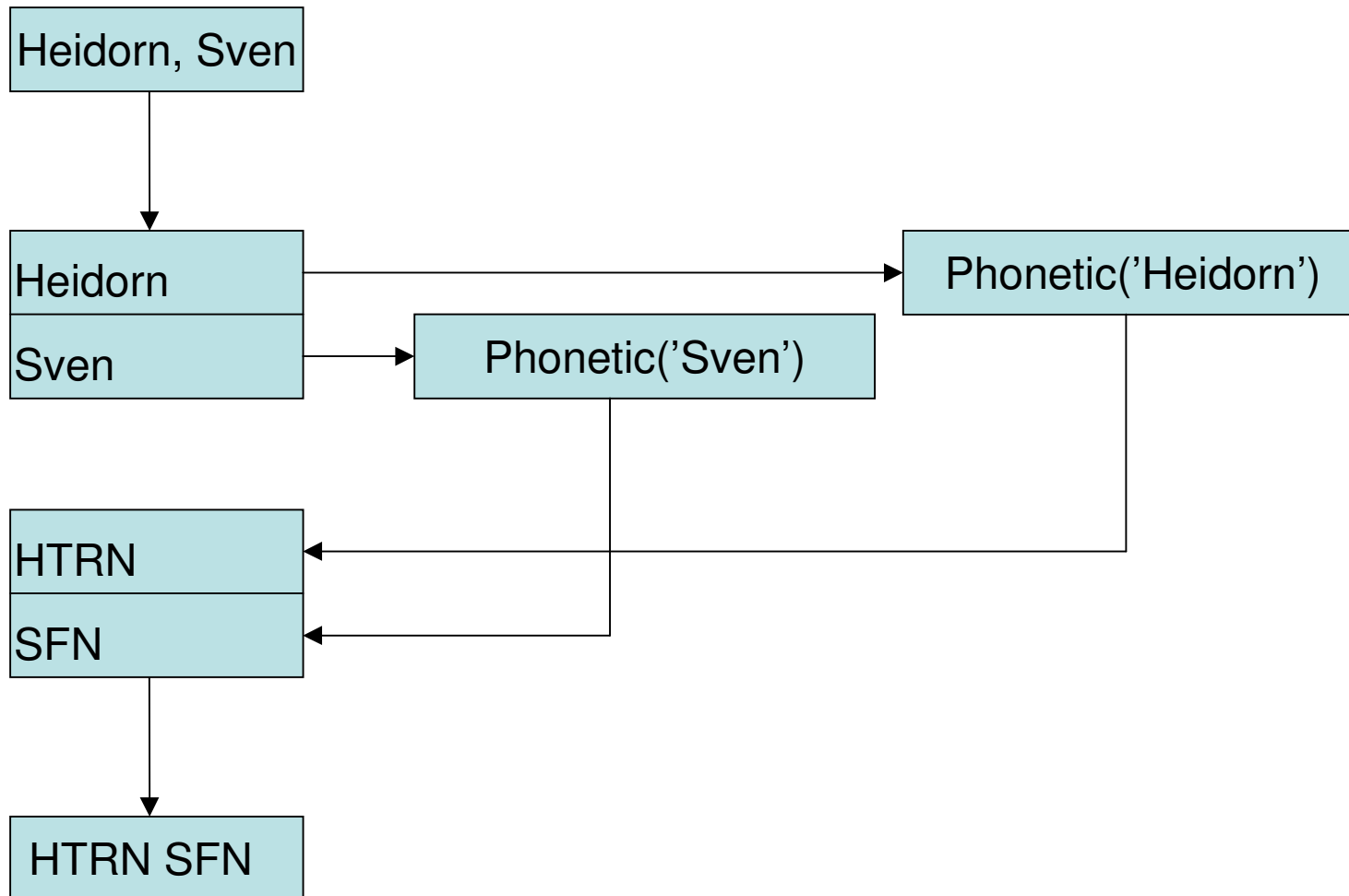
ANK%KL%

Fonetisk sökning

```
SELECT ID , NAME
       , NAME_PHONETIC
FROM your.userTable
WHERE NAME_PHONETIC LIKE 'ANK%KL%'
```

ID	NAME	NAME_PHONETIC
38009940	Anka, Kalle	ANKKL
16229163	ENGLUND ÖGGE, LINDA MARGARETA	ANKLNTKLNTMRKRT
12608295	WANGLER, JOHANNA ULRICA ELISE	ANKLRJHNLRLKS
38414389	INGMAN, ANDERS NICLAS	ANKMNNTRSCLKS
50304823	INKOMST, KALLE	ANKMSTKL
10202695	ENGBLOM, INGA ELISABETH LINNÉA	ANKPLMNKLSP0LN
43009935	ENGBERG, LARS ERIK	ANKPRKLRSRK
31244569	ANKARBERG, ULF ROGER	ANKRPRKLFRKR

Fonetisk sökning



Fonetisk sökning

- Bättre träffbild och därmed färre förvirrande resultat
- Likvärdigt med separata kolumner

```
SELECT ID , NAME
       , NAME_PHONETIC
FROM your.userTable
WHERE NAME_PHONETIC LIKE 'ANK%KL%'
```

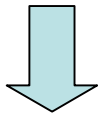
ID	NAME	NAME_PHONETIC
38009940	Anka, Kalle	ANK KL
38414389	INGMAN, ANDERS NICLAS	ANKMN ANTRS NKLS
50304823	INKOMST, KALLE	ANKMST KL

Fonetisk sökning

- z/Os
 - Double Metaphone har konverterats från C++ till Cobol
 - Cobol-koden har anpassats direkt för att kunna implementeras som en UDF (Phonetic)
 - Ytterligare en UDF (String_Phonetic) för att kunna hantera strängar med mer än ett namn

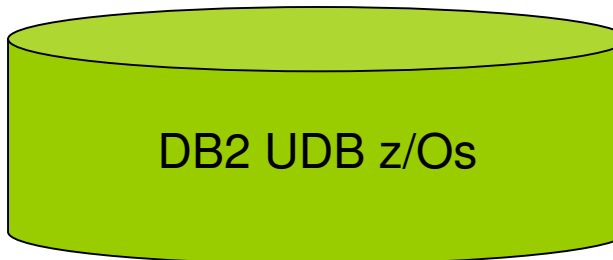
Fonetisk sökning

Phonetic UDF
Cobol

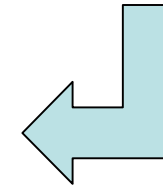


WLM

```
CREATE FUNCTION CSNFUN.PHONETIC (VARCHAR(128))
  RETURNS VARCHAR(128)
  SPECIFIC PHONETIC
  EXTERNAL NAME 'PHONETIC'
  LANGUAGE COBOL
  PARAMETER STYLE DB2SQL
  DETERMINISTIC
  RETURNS NULL ON NULL INPUT
  NO SQL
  NO EXTERNAL ACTION
  ALLOW PARALLEL
  WLM ENVIRONMENT DB22T9E1
  ASUTIME LIMIT 1024
  STAY RESIDENT YES
  PROGRAM TYPE SUB
  RUN OPTIONS 'TRAP(OFF),RPTOPTS(OFF),
H(,,ANY),STAC(,,ANY,),STO(,,,4K),BE(4K,,),LIBS(4K,,),ALL31(ON)';
```



DB2 UDB z/Os



Fonetisk sökning

- Utestående problem
 - LUW

```
SELECT ID,  
       NAME,  
       NAME_PHONETIC  
FROM your.userTable  
WHERE NAMN_PHONETIC LIKE RTRIM(RTRIM(PHONETIC('Anka')) CONCAT '%' CONCAT  
                               RTRIM(PHONETIC('Kalle')) CONCAT '%')
```

ID	NAME	NAME_PHONETIC
38009940	Anka, Kalle	ANK KL
38414389	INGMAN, ANDERS NICLAS	ANKMN ANTRS NKLS
50304823	INKOMST, KALLE	ANKMST KL

Fonetisk sökning

- Utestående problem
 - z/Os

```
SELECT ID,  
       NAME,  
       NAME_PHONETIC  
FROM your.userTable  
WHERE NAME_PHONETIC LIKE RTRIM(RTRIM(PHONETIC('Anka')) CONCAT '%' CONCAT  
                                RTRIM(PHONETIC('Kalle')) CONCAT '%')
```

```
DSNT408I  SQLCODE = -132, ERROR: AN OPERAND OF LIKE IS NOT VALID  
DSNT418I  SQLSTATE = 42824 SQLSTATE RETURN CODE  
DSNT415I  SQLERRP = DSNHSM2P SQL PROCEDURE DETECTING ERROR  
DSNT416I  SQLERRD = 0 0 0 -1 295 0 SQL DIAGNOSTIC INFORMATION  
DSNT416I  SQLERRD = X'00000000' X'00000000' X'00000000' X'FFFFFFFF'  
X'00000127' X'00000000' SQL DIAGNOSTIC INFORMATION
```

Fonetisk sökning

- The expression can be specified by any one of the following:
 - A constant
 - A special register
 - A host variable (including a LOB locator variable)
 - A scalar function whose arguments are any of the above (though nested function invocations cannot be used)
 - A CAST specification whose arguments are any of the above
 - An expression that concatenates (using CONCAT or ||) any of the above

Fonetisk sökning

- Utestående problem
 - z/Os

```
SELECT ID,  
       NAME,  
       NAME_PHONETIC  
FROM your.userTable  
WHERE NAME_PHONETIC LIKE PHONETIC('Anka') CONCAT '%'
```

```
DSNT408I SQLCODE = -132, ERROR: AN OPERAND OF IS NOT VALID  
DSNT418I SQLSTATE = 42824 SQLSTATE RETURN CODE  
DSNT415I SQLERRP = DSNXOW2D SQL PROCEDURE DETECTING ERROR  
DSNT416I SQLERRD = -202 0 0 -1 0 0 SQL DIAGNOSTIC INFORMATION  
DSNT416I SQLERRD = X'FFFFFFF36' X'00000000' X'00000000' X'FFFFFFF'  
X'00000000' X'00000000' SQL DIAGNOSTIC INFORMATION
```

– ????????

Fonetisk sökning

- Mer info

- <http://aspell.net/metaphone/>
- <http://jakarta.apache.org/commons/codecs/>
- http://www.cling.gu.se/theses/2002/cl8uglad_cl8mkarl.pdf
- <http://www.creativyst.com/Doc/Articles/SoundEx1/SoundEx1.htm>
- + otaliga websidor med artiklar och exempel på implementationer

Fonetisk sökning

- Frågor