@cczarski folgen

ORACLE

Mehr Ergebnisse:
Linguistische und Ähnlichkeitssuche mit SQL

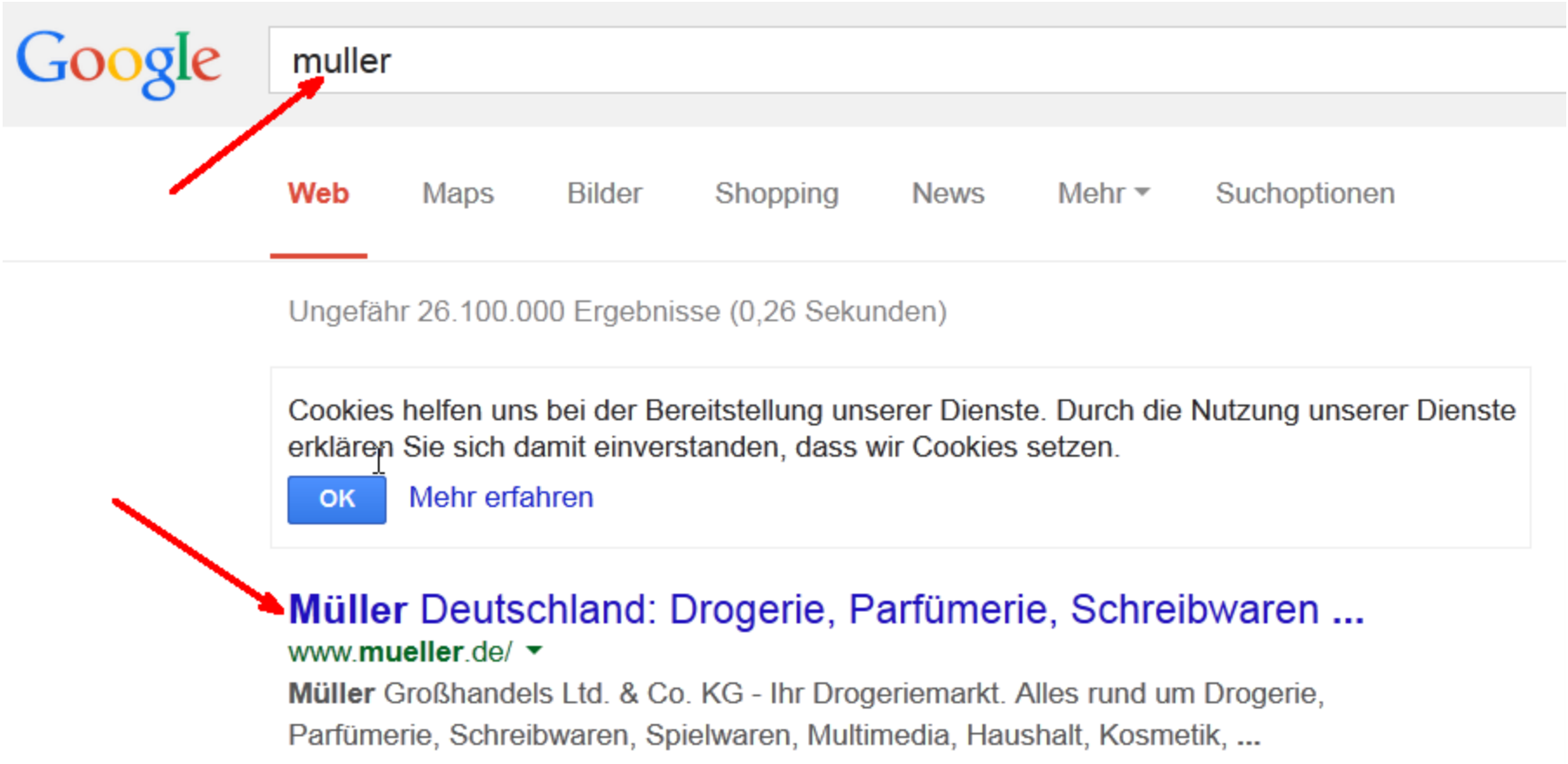
Carsten Czarski
ORACLE Deutschland B.V. & Co KG

Suche im DWH

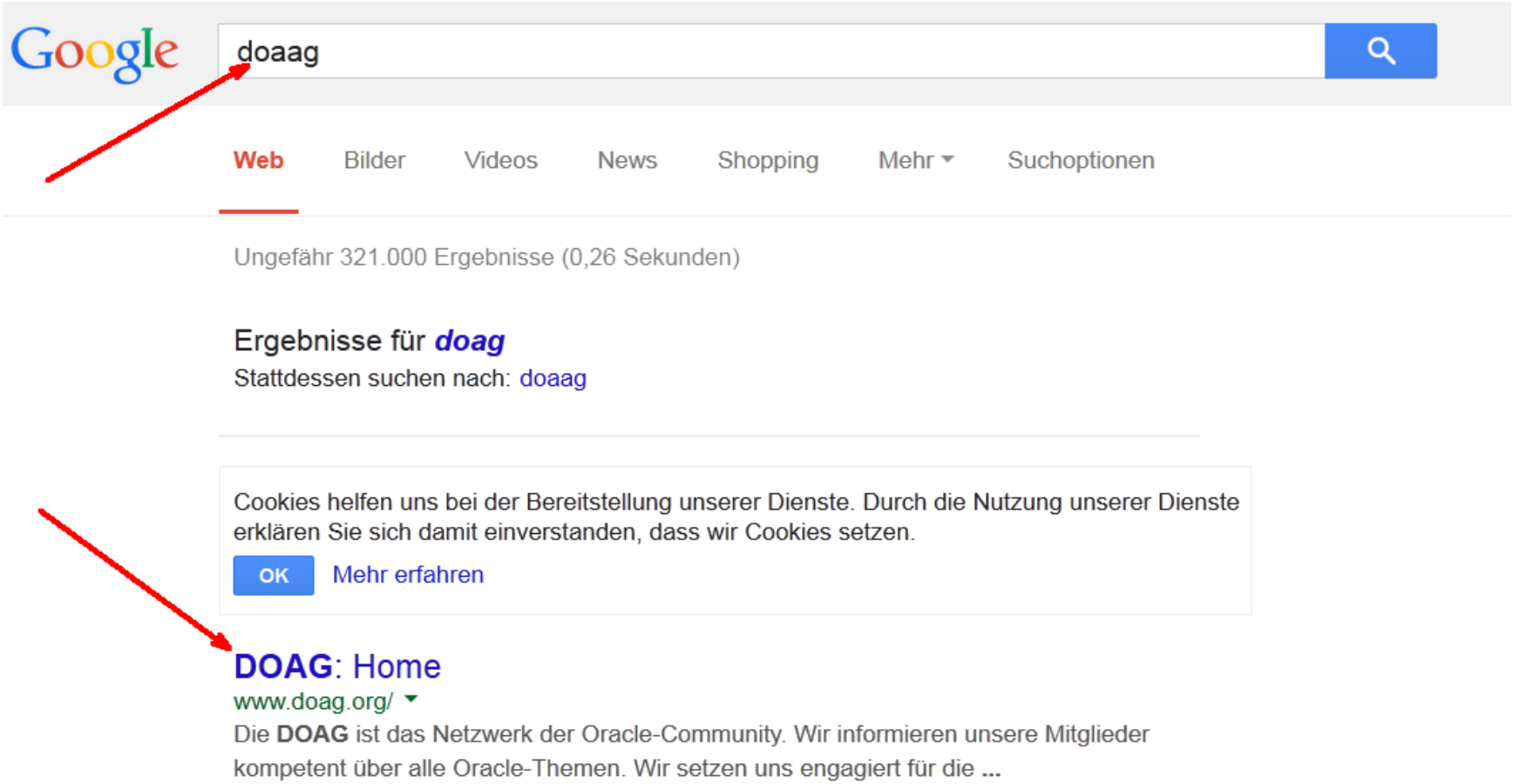
Was wünscht sich der Anwender ...?

Suche im DWH

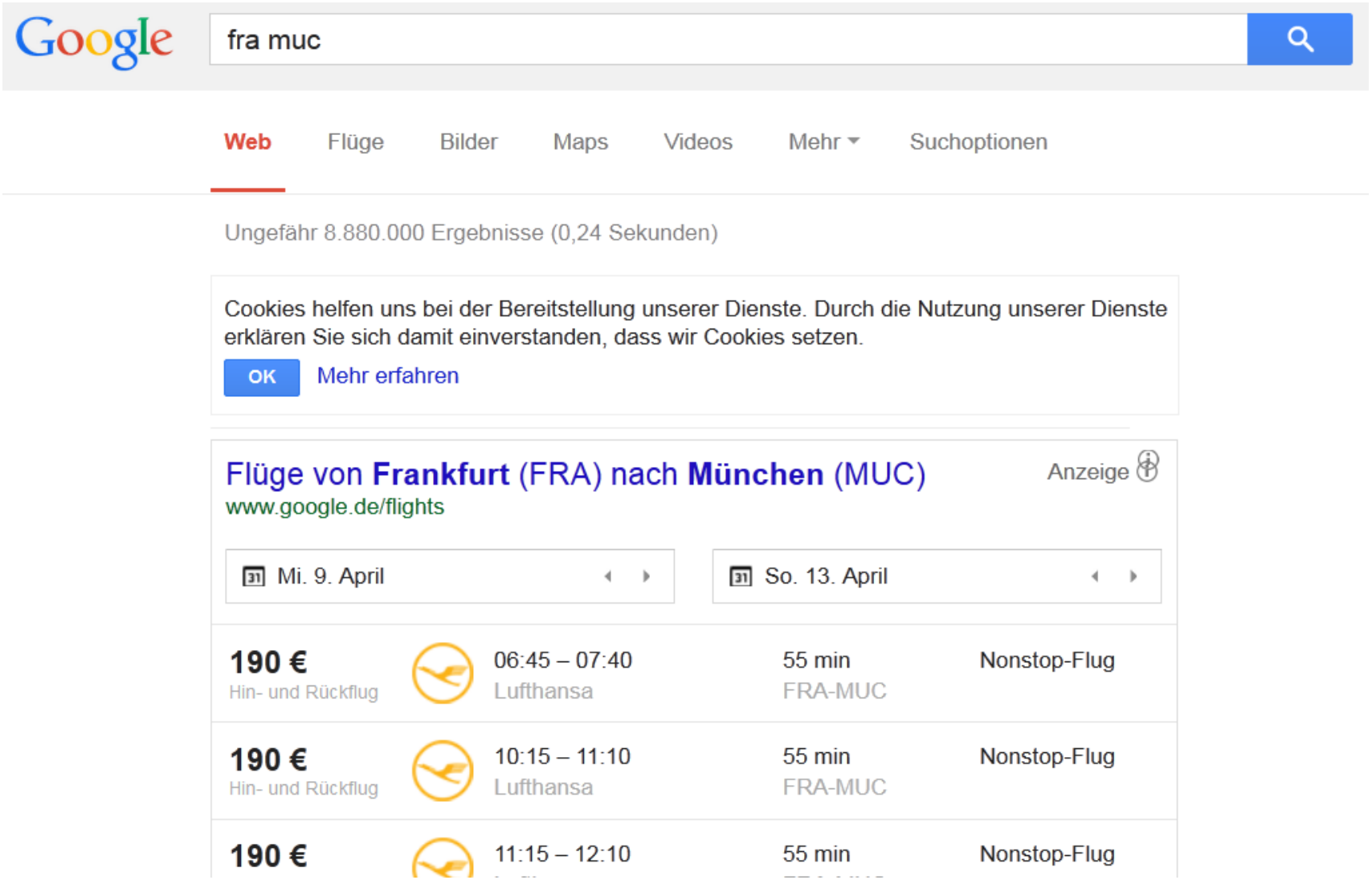
Suche im DWH


Google
muller
Web
Maps
Bilder
Shopping
News
Mehr
Suchoptionen
Ungefähr 26.100.000 Ergebnisse (0,26 Sekunden)
Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
OK
Mehr erfahren
Müller Deutschland: Drogerie, Parfümerie, Schreibwaren ...
www.mueller.de/
Müller Großhandels Ltd. & Co. KG - Ihr Drogeriemarkt. Alles rund um Drogerie, Parfümerie, Schreibwaren, Spielwaren, Multimedia, Haushalt, Kosmetik, ...

Suche im DWH


Google
doaag
Web
Bilder
Videos
News
Shopping
Mehr
Suchoptionen
Ungefähr 321.000 Ergebnisse (0,26 Sekunden)
Ergebnisse für doag
Stattdessen suchen nach: doaag
Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
OK
Mehr erfahren
DOAG: Home
www.doag.org/
Die DOAG ist das Netzwerk der Oracle-Community. Wir informieren unsere Mitglieder kompetent über alle Oracle-Themen. Wir setzen uns engagiert für die ...

Suche im DWH


Google
fra muc
Web
Flüge
Bilder
Maps
Videos
Mehr
Suchoptionen
Ungefähr 8.880.000 Ergebnisse (0,24 Sekunden)
Cookies helfen uns bei der Bereitstellung unserer Dienste. Durch die Nutzung unserer Dienste erklären Sie sich damit einverstanden, dass wir Cookies setzen.
OK
Mehr erfahren
Flüge von Frankfurt (FRA) nach München (MUC)
Anzeige
www.google.de/flights
Mi. 9. April
So. 13. April
190 €
Hin- und Rückflug
06:45 – 07:40
Lufthansa
55 min
FRA-MUC
Nonstop-Flug
190 €
Hin- und Rückflug
10:15 – 11:10
Lufthansa
55 min
FRA-MUC
Nonstop-Flug
190 €
11:15 – 12:10
55 min
Nonstop-Flug

Suchfunktion: Anforderungen

- Case-Insensitive Suche
- Umlaut-Insensitive Suche
- Fehlertolerante Suche
- Suche über *mehrere* Attribute

Case-Insensitive Suche

- SQL Funktionen UPPER und LOWER

```
select cust_first_name, cust_last_name, cust_city from
sh.customers
where upper(cust_last_name) = upper(:EINGABE);
```

- Performance: Funktionsbasierter Index

```
create index fix_upper_cust_lastname
on customers(upper(cust_last_name));

Index wurde erstellt.
```

Suchfunktion: Anforderungen

- ☑ Case-Insensitive Suche
- ☐ Umlaut-Insensitive Suche
- ☐ Fehlertolerante Suche
- ☐ Suche über möglichst *alle* Attribute

Linguistische Unterstützung in SQL

- Berücksichtigung spezieller Zeichen bei ...
  - Sortierung (ORDER BY)
  - Filter (WHERE)
  - Index-Lookup
- Gesteuert durch Session-Parameter
  - NLS_COMP = BINARY | LINGUISTIC
  - NLS_SORT stellt die konkrete Sprache oder BINARY ein - Suffix beachten

Liguistische Unterstützung in SQL

- Suffix für NLS_SORT
  - _CI: Ignorieren von Groß-/Kleinschreibung
  - _AI: Ignorieren von diakritischen Zeichen (è, á, ü)

- Beispiele

```
dbms_session.set_nls('NLS_SORT','BINARY_AI');

ALTER SESSION SET NLS_SORT=BINARY_AI;

select * from sortierung
order by nlssort(text,'NLS_SORT=BINARY_AI');
```

Liguistische Unterstützung in SQL

- Konkrete Anwendung

```
ALTER SESSION SET NLS_SORT=BINARY_AI;
ALTER SESSION SET NLS_COMP=LINGUISTIC;

select cust_first_name, cust_last_name, cust_city
from sh.customers
where cust_last_name = 'bäèr';

CUST_FIRST_NAME      CUST_LAST_NAME       CUST_CITY
-------------------- -------------------- --------------
Bryan                Baer                 Walsall
Bryan                Baer                 Evinston
Bryan                Baer                 Noma
Bryan                Baer                 Montara
```

Linguistischer Index

- Wichtig für Abfrageperformance

```
create index fidx_nls_customers
on customers (nlssort(cust_last_name, 'NLS_SORT=BINARY_AI'));

Index wurde erstellt.
```

- Ausführungsplan danach ...

```
|  0 | SELECT STATEMENT                    |                     | 555 |
|  1 | TABLE ACCESS BY INDEX ROWID BATCHED| CUSTOMERS           | 555 |
|* 2 | INDEX RANGE SCAN                    | FIDX_NLS_CUSTOMERS | 222 |
-------------------------------------------------------------------------

Predicate Information (identified by operation id):
---------------------------------------------------
2 - access(NLSSORT("CUST_LAST_NAME",'nls_sort=''BINARY_AI''')=HEXTORAW
('6261657200'))
```

http://sql-plsql-de.blogspot.co.uk/2007/11/case-und-umlaut-insensitive-suche-und.html

Was passiert da eigentlich ...?

- SQL-Funktion NLSSORT
  - Umwandlung einer Zeichenkette in einen "Schlüssel zum linguistischen Vergleich"
  - Rückgabe als RAW-Datentyp
  - Sprache als Parameter übergeben oder Default aus Session-Parameter NLS_SORT
- Spezialfall NLS_SORT=BINARY
  - Umwandlung der Zeichenkette in RAW-Datentyp
  - Vorherige Anwendung des _AI oder _CI Suffix

Was passiert da eigentlich ...?

- Ausgabe von NLSSORT

```
select nlssort('AaÄäUuÜüÖöOoßêÊéè','NLS_SORT=BINARY_AI') as RESULT from dual;

RESULT
----------------------------------------
61616161757575756F6F6F6F73736565656500

1 Zeile wurde ausgewählt.

select utl_raw.cast_to_varchar2(
  nlssort('AaÄäUuÜüÖöOoßêÊéè','NLS_SORT=BINARY_AI')
) as RESULT from dual;

RESULT
--------------------------------------------------------------------------------
aaaauuuuooooosseeee

1 Zeile wurde ausgewählt.
```

RAW-Datentyp

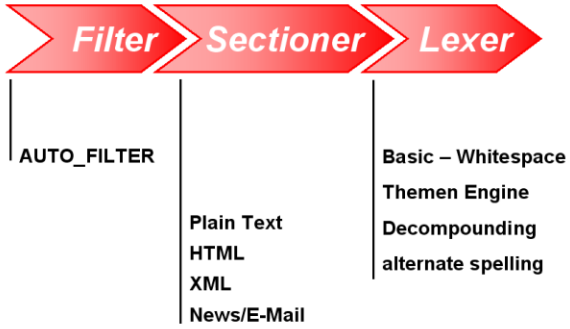
http://sql-plsql-de.blogspot.co.uk/2009/12/umlaute-normalisieren-nlssort.html

Suchfunktion: Anforderungen

- ☑ Case-Insensitive Suche
- ☑ Umlaut-Insensitive Suche
- ☐ Fehlertolerante Suche
- ☐ Suche über möglichst *alle* Attribute

Fehlertolerante Suche
Ähnlichkeitssuche mit Oracle TEXT

- Enthält viele linguistische Funktionen (mehr als SQL)
- Eigentlich für Suche in *Dokumenten* vorgesehen
- Nicht im DWH anwendbar, oder ...?



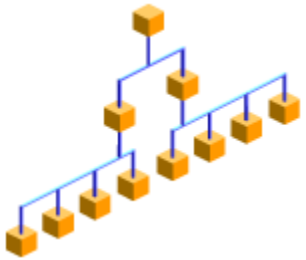


Tabelle mit Dokumenten

Index-Engine

Volltextindex

Die Indizierung

- Unterstützung aller gängigen Datentypen
  - XMLTYPE, VARCHAR2, CLOB, BLOB, Database Filesystem, SecureFile
  - Filter bei Binärdaten möglich

```
SQL> create index idx_textindex
  2  on dokument_tab (dokument)
  3  indextype is CTXSYS.CONTEXT
  4  /

Index created.
```

Funktionen der Volltextrecherche

- Abfrage mit SQL
  - CONTAINS-Funktion
  - Kombinierbar mit relationalen Abfragen
- Relevanz-Ranking anhand Wort-Häufigkeiten
  - SCORE()-Funktion
- Ergebnis-Aufbereitung
  - Highlighting
  - "Keyword-in-Context"

```
select score(1), dokument
from dokument_tab
where CONTAINS(dokument, 'Software AND Oracle')>0
/
```

Abfragemöglichkeiten

- Exakte Wort/Phrasensuche
  ... where contains(text, 'Hund')>0
- Logische Kombinationen
  ... where contains(text, 'Hund AND Katze') >0
- Wildcard-Suche
  ... where contains(text, 'Hu%d AND Kat_e') >0
- Namenssuche
  ... where contains(text, 'NDATA(name, Hunt)') >0
- Fuzzy matching
  ... where contains(text, '?Hunt') >0
- Multilinguale Stammsuche
  ... where contains(text, '$läuft') >0

Abfragemöglichkeiten

- NEAR-Operator
  … where contains(text, 'near(Hund, Katze), 4') >0
- Suche in Sektionen, Sätzen und Paragraphen (XML)
  … where contains(text, 'Hund WITHIN TITEL') >0
  Score-bezogene Funktionen
  … where contains(text, 'Hund MINUS Katze)' >0
  … where contains(text, 'Hund OR Katze*3)' >0
- Score-bezogene Operationen auf Ergebnislisten
  … where contains(text, 'Hund')>10
- ISO 2788 konformer Thesaurus
  … where contains(text, 'SYN(Hund,[thes]' >0)
- Soundex
  … where contains(text, '!Smythe') >0

Progressive Relaxation

- Erweiterung der Textquery
  - ... schrittweise ...
  - ... bis gewünschte Trefferzahl erreicht ist

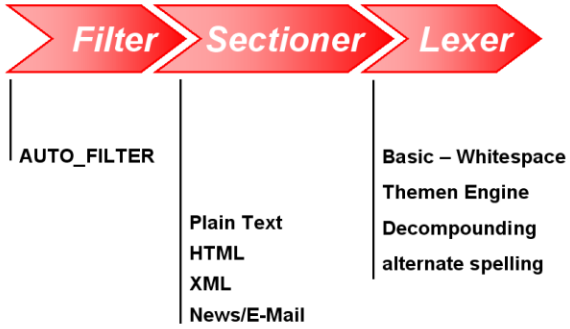
```
select score(1), title from test_table
where contains (
  dokument,
  '<query>
     <textquery>
       <progression>
         <seq>Hund and Katze</seq>
         <seq>Hund or Katze</seq>
         <seq>Hund? or Katze?</seq>
         :
and rownum <= :1
```

Oracle TEXT im Data Warehouse

- Anforderungen
  - Ähnlichkeitssuche
  - Einfache Suche über viele Attribute
- Ansatz
  - Volltextindex auf Tabelle im Data Warehouse
  - Nutzung eines *USER_DATASTORE*
  - Teil des Ladeprozesses

Oracle TEXT auf "normale" Tabellen
Was ist ein USER_DATASTORE?

- Normalfall (wie bei jedem Index)
  Inhalte der Indexspalte werden verarbeitet



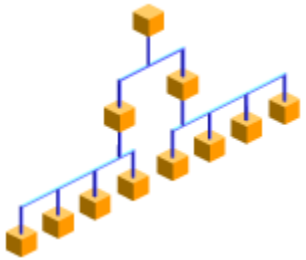


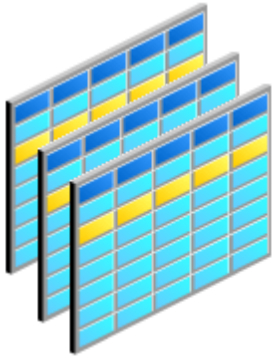
Tabelle mit Dokumenten

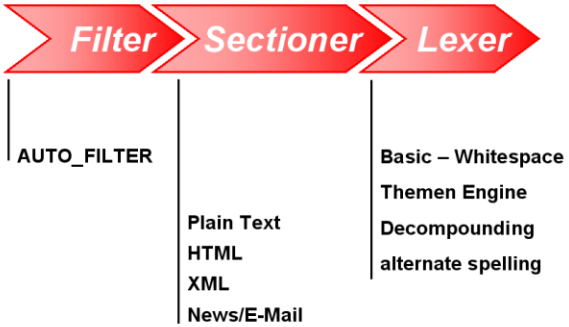
Index-Engine

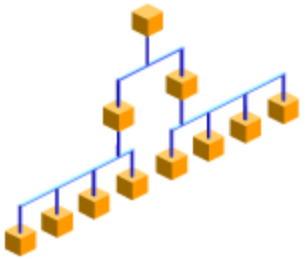
Volltextindex

Oracle TEXT auf "normale" Tabellen
Was ist ein USER_DATASTORE?

- USER_DATASTORE
  PL/SQL-Prozedur liefert die zu indizierenden Daten







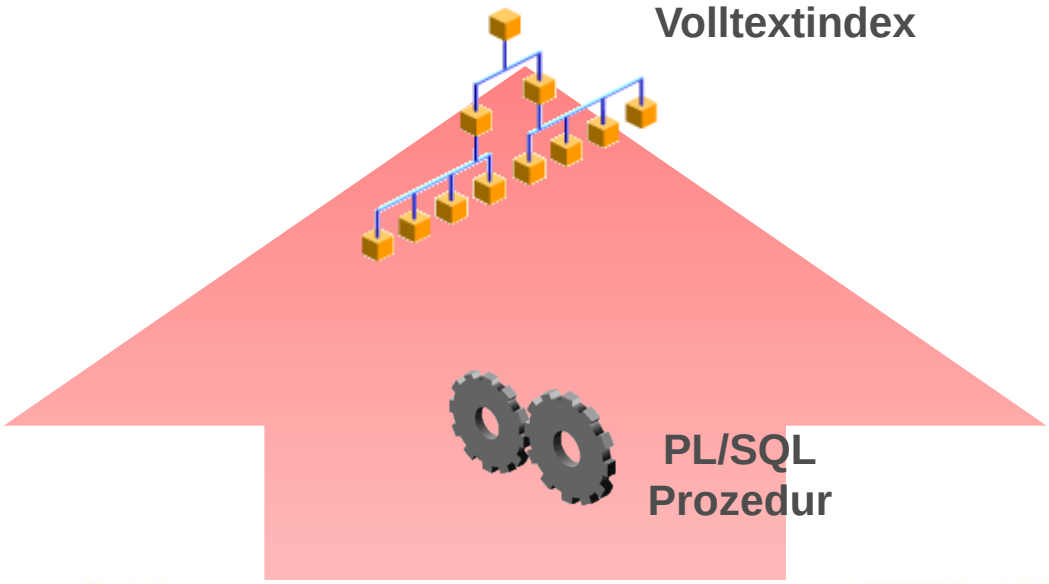
Tabelle(n)

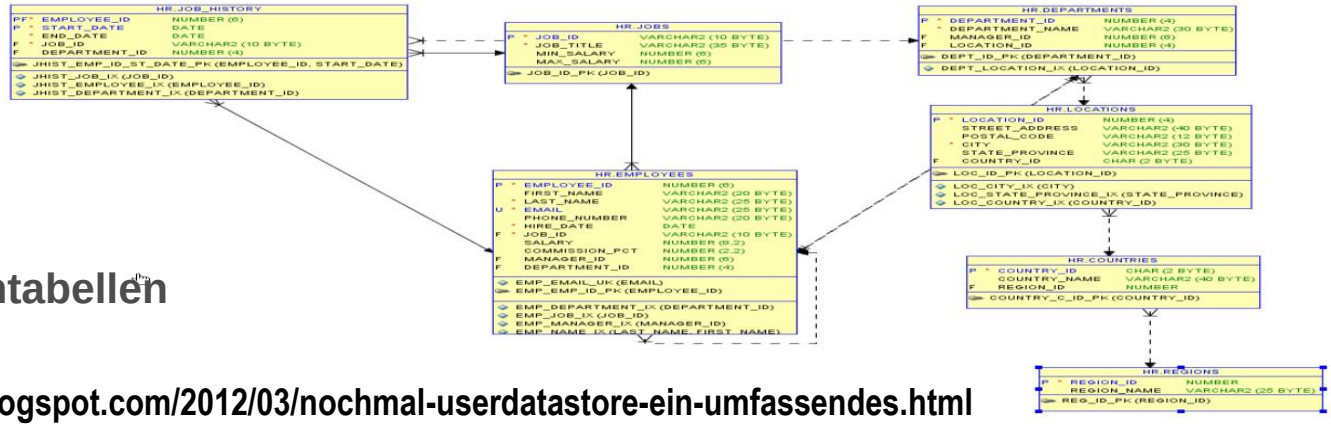
PL/SQL-Prozedur

Index-Engine

Volltextindex

Oracle TEXT auf "normale" Tabellen
Auch komplexe Datenstrukturen





Datentabellen

http://oracle-text-de.blogspot.com/2012/03/nochmal-userdatastore-ein-umfassendes.html

TEXT Index für "normale" Tabellen
Vorgehensweise

1. Suchtabelle erstellen
2. PL/SQL Prozedur für User Datastore erstellen
   - INPUT: **ROWID**
   - OUTPUT: VARCHAR2| CLOB | BLOB
3. Prozedur im Oracle TEXT Dictionary registrieren
4. Oracle TEXT Indexparameter einstellen
5. Oracle TEXT Index erstellen
6. Abfragen

Abfragebeispiele

```
Query: Oxxfort within (CITY) and SDATA(HIRE_DATE >= '2008-01-01')

EMPLOYEE_ID FIRST_NAME           LAST_NAME           CITY
----------- -------------------- ------------------- --------------------------
        149 Eleni                Zlotkey             Oxford
        164 Mattea               Marvins             Oxford
        165 David                Lee                 Oxford
        166 Sundar               Ande                Oxford
          : :                    :                   :

7 Zeilen ausgewählt.


Query: ?Accountant

EMPLOYEE_ID FIRST_NAME           LAST_NAME           JOB_TITLE           CITY
----------- -------------------- ------------------- ------------------- ----------
        205 Shelley              Higgins             Accounting Manager  Seattle
        206 William              Gietz               Public Accountant   Seattle
        109 Daniel               Faviet              Accountant          Seattle
        110 John                 Chen                Accountant          Seattle
        : :                      :                   :                   :

7 Zeilen ausgewählt.
```

Wartung eines Oracle TEXT Index
Oracle TEXT und DML

- Besonderheit ...
  - Idealfall: In-Place-Pflege
  - Aber: Zu teuer – Indexstrukturen sind komplex
- Daher:
  - INSERT: PENDING-Tabelle
  - DELETE: Negativliste
  - UPDATE: DELETE und INSERT
- Zusätzliche Aufgaben:
  - **Index-Synchronisierung**
  - **Index-Optimierungen**

Suchfunktion: Anforderungen

- ☑ Case-Insensitive Suche
- ☑ Umlaut-Insensitive Suche
- ☑ Fehlertolerante Suche
- ☑ Suche über möglichst *alle* Attribute

Weitere Informationen

- Deutschsprachiges Blog zu SQL und PL/SQL
  http://sql-plsql-de.blogspot.com
- Deutschsprachiges Blog zu Oracle TEXT
  http://oracle-text-de.blogspot.com

- Handbücher (Dokumentation)
  http://docs.oracle.com
  - Oracle SQL Language Reference
  - Oracle TEXT Application Developers Guide
  - Oracle TEXT Reference

Carsten.Czarski@oracle.com

http://tinyurl.com/apexcommunity

http://sql-plsql-de.blogspot.com
http://oracle-text-de.blogspot.com
http://oracle-spatial.blogspot.com

http://plsqlexecoscomm.sourceforge.net
http://plsqlmailclient.sourceforge.net

Twitter: @cczarski @oraclebudb