performing databases

Your reliability. Our concern.

performing databases
Your reliability. Our concern.

Oracle Text: AllesTextOderWas?

Benedikt Nahlovsky

Performing Databases GmbH
Mitterteich

Referent

Benedikt Nahlovsky

Datenbankspezialist & Performance-Firefighter

Fachliche Schwerpunkte:

- Performanceoptimierung / Tuning
- Cluster und Replikation
- Upgrade / Migration

Kontakt:
benedikt.nahlovsky@performing-db.com

Weblog:
http://oradbn.wordpress.com

Performing Databases

performing databases
Your reliability. Our concern.

Spezialisten für Datenbanktechnik

- Konzeptberatung
- Architektur- und Systemplanung
- Lizenzierung
- Realisierung und Troubleshooting

Kontakt:
Performing Databases GmbH
Wiesauer Straße 27
95666 Mitterteich

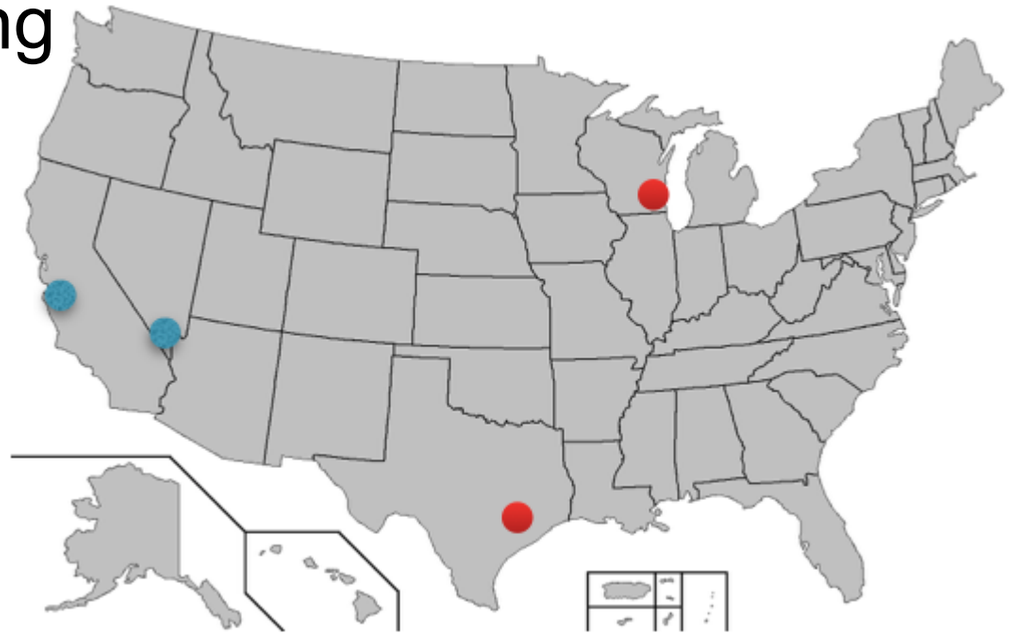
Web: http://www.performing-databases.com
Mail: info@performing-db.com
Twitter: @PerformingDB

International im Einsatz

performing databases
Your reliability. Our concern.

ORACLE Gold Partner

- Planung
- Lizenzierung
- Umsetzung
- Tuning
- Troubleshooting
- Wartung
- Upgrade
- Migration



Oracle Text

performing databases
Your reliability. Our concern.

- seit Oracle Version 7 die Möglichkeit zur Volltextsuche
- ab Version 8 in kostenpflichtiges Modul interMedia integriert
- seit Version 9i fester Bestandteil der Datenbank
- in allen Versionen kostenlos (XE, SE, SE2, EE)
- zur sofortigen Verwendung sind nur wenige Schritte notwendig

Installation

- Installation notwendig, um die Oracle Text PL/SQL Packages zu verwenden
- Funktionen im Schema CTXSYS enthalten

```
SQL> conn / as sysdba
Connected.
SQL> @?/ctx/admin/catctx.sql ctxsys SYSAUX TEMP NOLOCK
...creating user CTXSYS
```

- User benötigen die Rolle CTXAPP
- Standard Spracheinstellung hinterlegen

```
SQL> conn ctxsys/ctxsys
Connected.
SQL> @?/ctx/admin/defaults/dr0defin.sql "GERMAN";
```

Überprüfung

```
SQL> select comp_name, status, version from dba_registry where
comp_id='CONTEXT';
COMP_NAME                      STATUS       VERSION
------------------------------ ------------ -----------
Oracle Text                    VALID        11.2.0.4.0
```

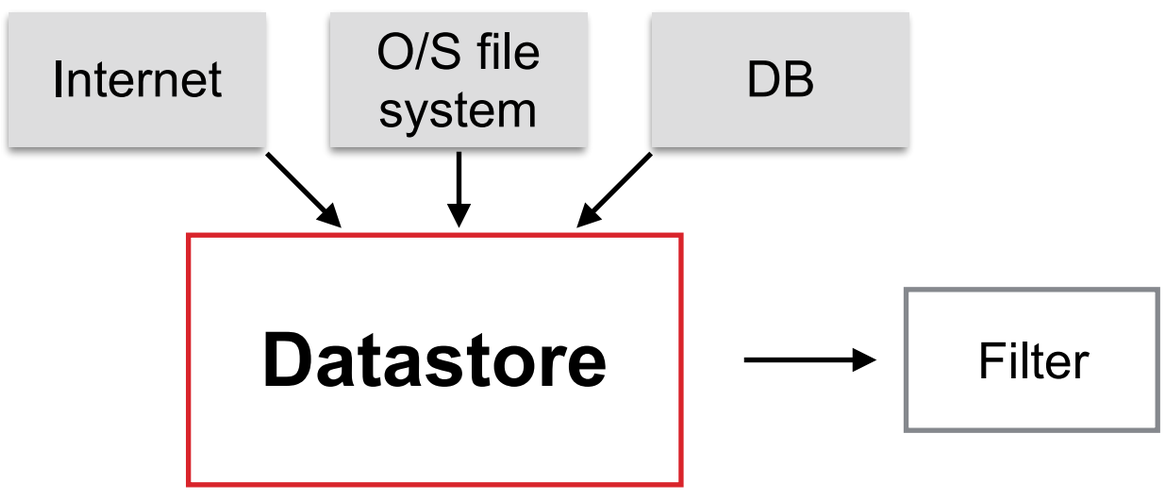
```
SQL> grant ctxapp to perfdbtext;
Grant succeeded.
```

—> Oracle Text kann mit dem Schema perfdbtext nun verwendet werden

Wie funktioniert Oracle Text?

performing databases
Your reliability. Our concern.

1. Dokumente werden in sogenannte Datastores eingelesen
   - DIRECT_DATASTORE (CLOB, VARCHAR2, etc.)
   - FILE_DATASTORE (Filesystem am DB Server)
   - URL_DATASTORE (Inter- bzw. Intranet)
   - USER_DATASTORE (selbst definierte Prozedur)



Wie funktioniert Oracle Text?

performing databases
Your reliability. Our concern.

2. Filter

- Umwandlung von Binärdokumenten in Text oder HTML
- nur bei Word- oder PDF-Dokumenten notwendig
- Text-, HTML- und XML-Dateien ausgeschlossen
- Oracle erkennt über 150 Formate automatisch

Wie funktioniert Oracle Text?

performing databases
Your reliability. Our concern.

3. Sectioner
   - teilt HTML- oder XML-Dokumente anhand von Tags in einzelne Abschnitte auf
   - HTML: <H1>…</H1>
   - XML: <Produktbeschreibung>…</Produktbeschreibung>

Wie funktioniert Oracle Text?

performing databases
Your reliability. Our concern.

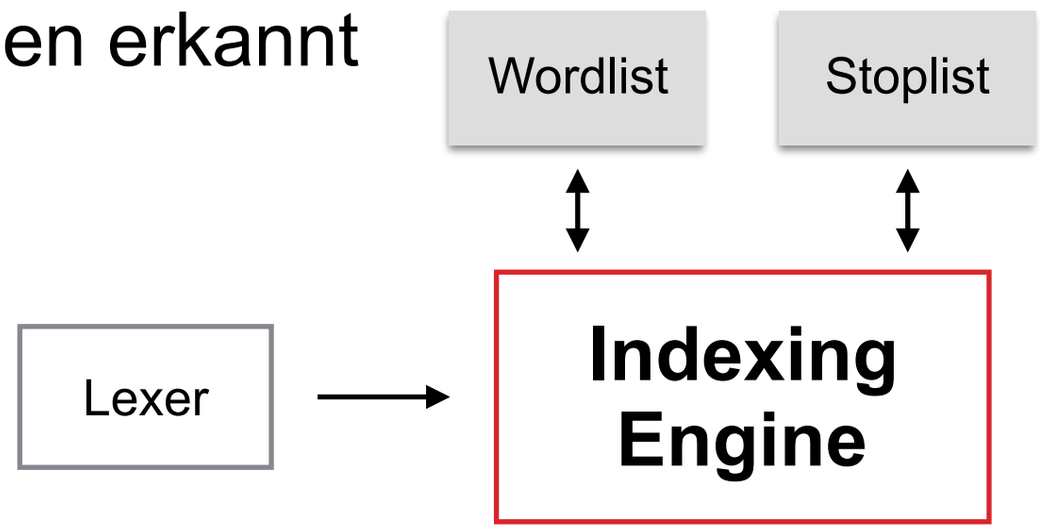
4. Lexer

- extrahiert relevante Wörter aus dem Text (Tokens)
- Interpunktions- und Sonderzeichen werden entfernt
- Trennzeichen Definition
- Case Sensitive / Case Insensitive
- Aufteilung zusammengesetzter Worte

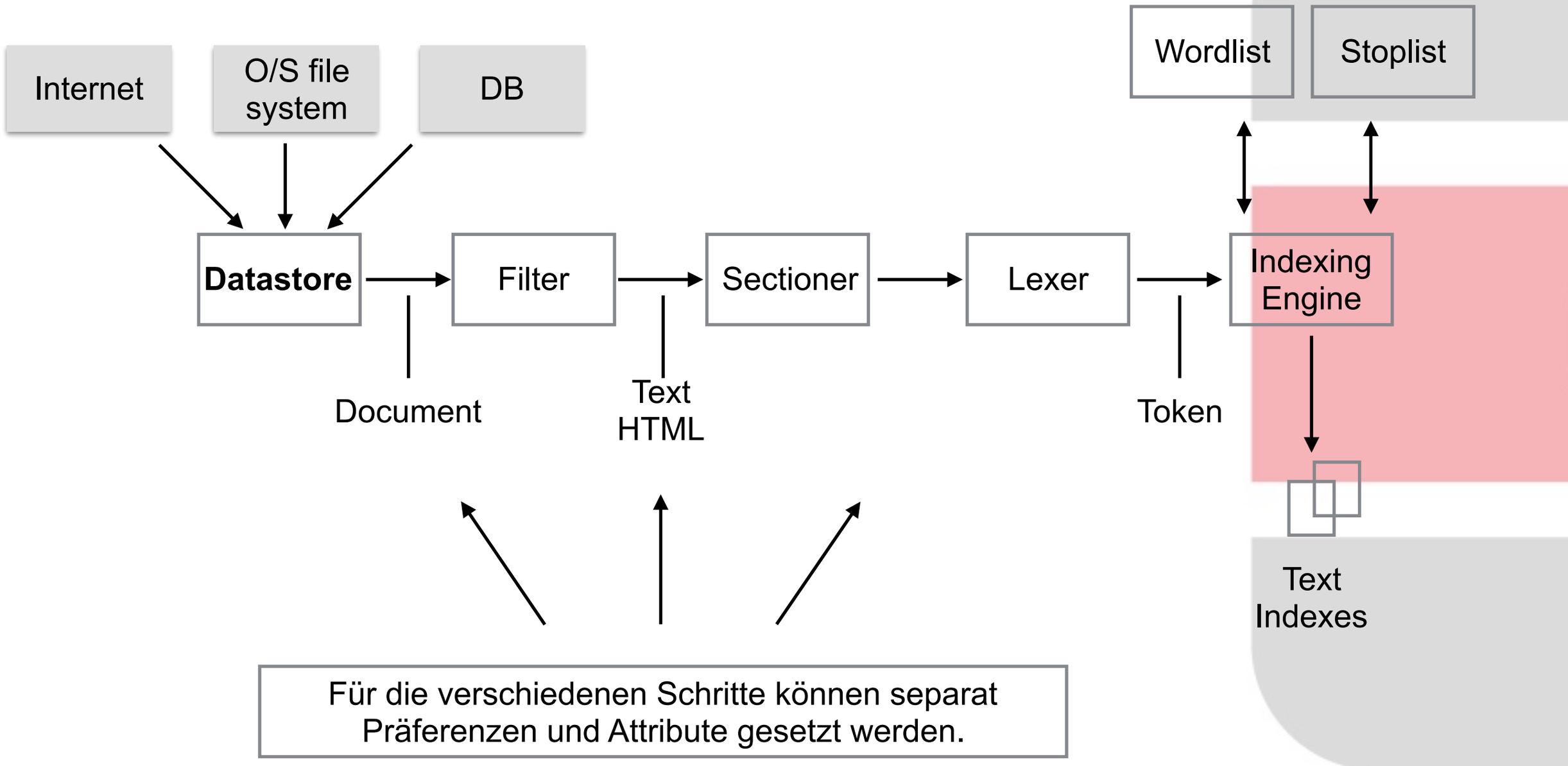
Wie funktioniert Oracle Text?

performing databases
Your reliability. Our concern.

5. Indexing Engine
   - aus den gesammelten Worten wird invertierter Index erzeugt
   - Oracle Text Stoplists werden nicht indiziert
   - Wordlist Einstellungen anhand grammatikalischer Regeln
   - Beugungsformen des gesuchten Wortes mit ähnlichem Stamm werden erkannt



Wie funktioniert Oracle Text?



Stoplist

performing databases
Your reliability. Our concern.

- Wörter werden nicht indiziert
- bei Abfragen werden diese nicht berücksichtigt
- CTX.DEFAULT_STOPLIST (abhängig von Spracheinstellung)
- Bindewörter, Präpositionen, Artikel betroffen
- Anwenderspezifische Stoplists können erstellt werden
- in der Tabelle CTX_STOPWORDS sind die Stopwords zu finden

  mit Zuordnung zu welcher Stoplist diese gehören
- in der Tabelle CTX_STOPLISTS die Listen mit Schemazuordnung

Stoplist anlegen

performing databases
Your reliability. Our concern.

- Stoplists sollten immer im entsprechenden Schema angelegt werden
- branchenspezifische Stopwords können hinterlegt werden

```
begin
  ctx_ddl.add_stopword(
    'AIRPORT_STOPLIST',
    'Abflug'
  );
end;
```

Ablage der Indexinformationen

performing databases
Your reliability. Our concern.

Ablage in DR$ Tabellen - Naming-Convention immer gleich

- Es beginnt mit dem Präfix DR
- Bei nicht partitionierten Indizes folgt ein „$“, bei partitionierten Indizes ein „#“
- Dann folgt der Name des Volltextindexes
- Bei einem partitionierten Index folgt auf den Namen eine Partitions-ID
- Es folgt ein „$“
- Abschließend folgt ein Suffix, welches die genaue Aufgabe der Tabelle bezeichnet

Syntax: DR$<NAME>$<SUFFIX>

$I: Token Tabelle

- eigentliche Index Tabelle
- speichert die Tokens (einzelne Wörter) welche sich durch die Zerlegung des Fließtextes ergeben haben ab (TOKEN_TEXT)
- Spalte TOKEN_INFO enthält die Informationen, in welchem Dokument, an welchen Stellen das jeweilige Token vorkommt

Syntax: DR$<NAME>$I

$R: Mapping von DOCID auf ROWID

- Oracle Text arbeitet intern nicht mit ROWIDs als Zeiger auf die Tabellenzeilen, sondern mit DOCIDs
- erste Zeile erhält DOCID 1, dann wird weitergezählt
- ein Volltext-Index ist eine invertierte Liste - mehrere Einträge können auf eine Tabellenzeile zeigen
- anhand der DOCID kann die ROWID herausgesucht werden

Syntax: DR$<NAME>$R

$K: Mapping von ROWID auf DOCID

- ist das Gegenstück zur $R-Tabelle
- anhand einer ROWID findet Sie die für den Textindex relevanten DOCIDs heraus
- wichtig bei DML Operationen

Syntax: DR$<NAME>$K

$N: Negativliste

- enthält gelöschte Dokumente
- Oracle Text arbeitete prinzipiell asynchron
- bei INSERT-Anweisung wird die ROWID in eine Pending - Tabelle eingetragen (SYNC-Operation notwendig)
- bei DELETE-Anweisung wird die DOCID ermittelt ($K-Tabelle) und trägt diese in die Negativliste ein
- bei SELECT-Operationen ermittelt Oracle Text die Treffermenge ganz normal und filtert dann die Einträge der Negativliste heraus

Syntax: DR$<NAME>$N

Index Types

Index Type	Application Type	Query Operator
CONTEXT	• Text besteht aus großen Dokumenten • Word, HTML, XML	CONTAINS
CTXCAT	• kleine Textbruchstücke • Artikelnamen, Preise, Beschreibungen	CATSEARCH
CTXRULE	• Dokumenten-klassifizierung • Anwendung auf Such-Strings	MATCHES

Beispiel Context Index

```
CREATE TABLE TEXTTAB
(
   ID NUMBER NOT NULL
, TEXT VARCHAR2(800)
, CONSTRAINT TEXTTAB_PK PRIMARY KEY
(
ID
)
ENABLE );
Table created.
SQL>

…rows inserted…

SQL> CREATE INDEX text_idx ON texttab(text) INDEXTYPE IS ctxsys.context;
Index created.
SQL>
```

Beispiel Context Index

performing databases
Your reliability. Our concern.

Syntax:

```
SELECT spaltenliste FROM tabelle WHERE CONTAINS(index_spalte,
'<suchbegriff>')>0;
```

Einfache Suche nach Wörtern, z.B.:

```
SQL> SELECT * FROM texttab WHERE CONTAINS(text, 'Landkreis') > 0;
  ID    TEXT
  ----------------------------------------------------------
  5     Der Landkreis Schwandorf liegt in Bayern.
  6     Der Landkreis Regensburg ist größer als Schwandorf.
  9     Mitterteich liegt im Landkreis Tirschenreuth.
SQL>
```

Beispiel Context Index

performing databases
Your reliability. Our concern.

Suche nach Wort-Kombinationen oder -Alternativen mit den Booleschen Operatoren „AND“ und „OR“

```
SQL> SELECT * FROM texttab WHERE CONTAINS(text, 'Landkreis AND Bayern') > 0;

  ID    TEXT
  ----------------------------------------------------------
  5     Der Landkreis Schwandorf liegt in Bayern.
SQL>


SQL> SELECT * FROM texttab WHERE CONTAINS(text, 'Landkreis OR Bayern') > 0;
  ID    TEXT
  ----------------------------------------------------------
  2     Die Landeshauptstadt von Bayern ist Munchen.
  5     Der Landkreis Schwandorf liegt in Bayern.
  6     Der Landkreis Regensburg ist größer als Schwandorf.
  9     Mitterteich liegt im Landkreis Tirschenreuth.
SQL>
```

Beispiel Context Index

Suche nach ähnlich geschriebenen Wörtern mit dem Operator „?“

```
SQL> SELECT * FROM texttab WHERE CONTAINS(text, '?Baiern') > 0;

  ID    TEXT
  ----------------------------------------------------------
  2     Die Landeshauptstadt von Bayern ist Munchen.
  5     Der Landkreis Schwandorf liegt in Bayern.
SQL>
```

Suche mit Wildcards: „%“ für kein oder beliebig viele Zeichen und „_“ für genau 1 Zeichen

```
SQL> SELECT * FROM texttab WHERE CONTAINS(text, '%dor_') > 0;

  ID    TEXT
  ----------------------------------------------------------
  5     Der Landkreis Schwandorf liegt in Bayern.
  6     Der Landkreis Regensburg ist größer als Schwandorf.
SQL>
```

Beispiel Context Index

Suche nach Ausdrücken, die denselben Wortstamm haben wie das Suchwort oder mit dem Suchwort zusammengesetzte Worte bilden, mit dem Operator „$“:

```
SQL> SELECT * FROM texttab WHERE CONTAINS(text, '$liegen') > 0;

  ID    TEXT
  ----------------------------------------------------------
  5     Der Landkreis Schwandorf liegt in Bayern.
  9     Mitterteich liegt im Landkreis Tirschenreuth.
SQL>
```

Index Sync

performing databases
Your reliability. Our concern.

1. Manuell

```
SQL> INSERT INTO texttab VALUES(11, 'New York ist eine Stadt an der
Ostseekuste in den USA');
1 row created.

SQL> SELECT * FROM texttab WHERE CONTAINS(text, 'USA') > 0;
no rows selected

SQL> exec ctx_ddl.sync_index('text_idx');
PL/SQL procedure successfully completed.

SQL> SELECT * FROM texttab WHERE CONTAINS(text, 'USA') > 0;

  ID    TEXT
  ----------------------------------------------------------------
  11    New York ist eine Stadt an der Ostkuste in den USA
SQL>
```

Index Sync

performing databases
Your reliability. Our concern.

2. Automatisch in regelmäßigen Intervallen

beim Anlegen (create index)

```
SQL> CREATE INDEX text_idx ON texttab(text) INDEXTYPE IS CTXSYS.CONTEXT
PARAMETERS ('SYNC (EVERY "TRUNC(sysdate)+1/24")');
Index created.
```

oder nachträglich (alter index)

```
SQL> ALTER INDEX text_idx REBUILD PARAMETERS (REPLACE METADATA 'SYNC (EVERY
"TRUNC(sysdate)+1/24")';
Index altered.
```

Index Sync

performing databases
Your reliability. Our concern.

3. Automatisch nach Commit (ab 10g - seltene DML Operationen)

beim Anlegen (create index)

```
SQL> CREATE INDEX text_id ON texttab(text) INDEXTYPE IS ctxsys.context
PARAMETERS ('SYNC (ON COMMIT)');
Index created.
```

oder nachträglich (alter index)

```
SQL> ALTER INDEX text_id REBUILD PARAMETERS ('REPLACE METADATA SYNC (ON
COMMIT)');
Index altered.
```

Index Sync

performing databases
Your reliability. Our concern.

4. Automatisch nach jeder Transaktion (ab 10g)

beim Anlegen (create index)

```
SQL> CREATE INDEX text_id ON texttab(text) INDEXTYPE IS ctxsys.context
PARAMETERS ('TRANSACTIONAL');
Index created.
```

oder nachträglich (alter index)

```
SQL> ALTER INDEX text_id REBUILD PARAMETERS ('REPLACE METADATA
TRANSACTIONAL)');
Index altered.
```

Index Optimierung

performing databases
Your reliability. Our concern.

CTX_DDL.OPTIMIZE_INDEX

- in regelmäßigen Abständen ausführen
- DBA_SCHEDULER_JOB

```
SQL> BEGIN
1 ctx_ddl.optimize_index(index_name => 'text_idx', optlevel => 'FULL');
2 END;
```

12c New Features

performing databases
Your reliability. Our concern.

- Erhöhung von MAX_INDEX_MEMORY für mehr OPTIMIZE und CREATE Performance
- Near Real Time Index: Häufige Index Syncs ohne Fragmentierung
- BIG_IO: Neue Indexarchitektur reduziert Index Lookups bei großen Indexfragmenten
- Pattern Stopclass: reguläre Ausdrücke für Tokens, die nicht indiziert werden sollen

Tipps

- „großer“ TEMP Tablespace
- Oracle Text Indizes sollten ständig optimiert und gewartet werden
- Index-Tabellen klein halten
- nutzen Sie Partitionierung (DR$-Tabellen werden aufgeteilt)

Fazit

- Oracle Text ist ein mächtiges und oftmals unterschätztes Werkzeug
- feste Integration, keine zusätzlichen Kosten
- ständige Weiterentwicklung in den neuen Datenbank-Versionen

performing databases

Your reliability. Our concern.

Download Präsentation und Whitepaper

http://www.performing-databases.com

performing databases

Your reliability. Our concern.