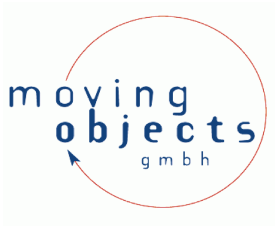
Zeichensatzkonvertierung in Oracle-DB

moving objects GmbH
Martin Busik
Hamburg - Mai 2003
www.moving-objects.de

moving objects GmbH



- Beratung
  - Anforderungserhebung
  - Geschäftprozessanalyse
  - Coaching
- Schulung
  - OOA/OOD, Java, J2EE, JSP, Servlets, Swing, XML
- Realisierung
  - J2EE-Applikationen
  - Datenbankapplikationen (ORACLE)
- www.moving-objects.de

Thema
Zeichensatzkonvertierung

In unterschiedlichen Zeichensätzen können identische Zeichen unterschiedlich kodiert sein. Bei einer Zeichensatzkonvertierung gibt es Probleme, wenn ein gegebenes Zeichen in einem der Zeichensätze nicht definiert ist.

Font vs. Zeichensatz

- Ein Zeichensatz besagt welche Zeichen vorgesehen sind und welche (numerische) Darstellung bzw. Code ein Zeichen hat. Z.B. ASCII oder Unicode
- Ein Font legt fest wie Zeichen (=numerischer Code) eines Zeichensatzes grafisch dargestellt werden
- Die (graphische) Darstellung eines einzelnen Zeichen wird Glyph genannt.
- D.h. Darstellung und Verarbeitung sind getrennt, auch wenn ein Zeichen nicht dargestellt wird, kann es verarbeitet werden (z.B. Die Windows System-Schriftart) und umgekehrt.

Zeichensätze in Oracle

- Client/Server System, auch wenn man mittels sqlplus auf der gleichen Maschine wie die Datenbank arbeitet!
  - Client und Server können mit unterschiedlichen Zeichensätzen arbeiten
  - Mit Database Links verbundene Datenbanken können unterschiedliche Zeichensätze verwenden
- In der Datenbank wird der Zeichensatz bei der Erstellung festgelegt (create database Statement)
- Jeder Client (sqlplus, oci usw.) kann pro Datenbanksession einen eigenen Zeichensatz verwenden.
  - Festlegung über die Umgebungsvariable NLS_LANG (bzw. Registry)
    `export NLS_LANG=german_germany.WE8ISO8859P1`
- Java verwendet intern Unicode (UCS2)
  - Weniger Probleme mit Zeichensatzumwandlungen

Wo findet Konvertierung statt

Anbindung	Wo?	NLS_LANG	Anmerkungen
OCI (sqlplus, perl)	OCI Treiber	Verwendet	
ODBC	OCI Treiber	Verwendet	Verwendet OCI
JDBC OCI (Fat)	OCI Treiber	Nein	Zwischenschritt UTF
JDBC Thin	Server	Nein	Zwischenschritt UTF
Java Server JDBC	Server	Nein	Zwischenschritt UTF

Häufig verwendete Zeichensätze

Zeichensatz	Oracle Bezeichnung	Codelänge	Umlaute	€-Zeichen
	US7ASCII	7 Bit	Nein	Nein
OEM, IBMPC	WE8PC850	8 Bit	0x8E – 0xE0	Nein
Latin1	WE8ISO8859P1	8 Bit	0xC4 – 0xFC	Nein
“ANSI”	MSWIN1252	8 Bit	0xC4 – 0xFC	Ja (0x80)
	WE8PC858	8 Bit	0x8E – 0xE0	Ja (0xD5)
Latin9, Latin0	WE8ISO8859P15	8 Bit	0xC4 – 0xFC	Ja (0xA4)
UTF-8	UTF8	8 bis 24 Bit*	Ja	Ja
UCS-2	AL16UTF16	16 Bit	Ja	Ja
Mac	WE8MACROMAN8	8 Bit	Ja	Nein

Demo

- Es werden zwei Datenbanken verwendet, Datenbank alpha (mit dem Zeichensatz WE8PC850 erstellt) und beta (mit dem Zeichensatz WE8ISO8859P1 erstellt)

Ergebnisse der Demo

- WE8ISO8859P1 ist nicht WE8MSWIN1252, wird aber häufig gleichgesetzt
- WE8ISO8859P1 ist der Oracle-default Zeichensatz bei Oracle 8i und Deutsch als Systemsprache (locale)
- Solange keine Zeichensatzkonvertierung durchgeführt wird, werden scheinbar alle Zeichen unterstützt
- Probleme werden erst bei EAI- oder Web-Projekten sichtbar

National Character Set

- Zweiter Zeichensatz in der Datenbank (Datentypen nchar, nvarchar, nclob)
  - flexibler, es werden mehr Zeichensätze unterstützt
- In sqlplus wird aber nur ein Zeichensatz verwendet!
  - sinnvoll nur bei direkten Aufrufen der oci-Funktionen
- Verwendung des „national character set“ führen zu Komplikationen:

```
select * from emp where ename = N'Müller';
```

Möglichkeiten der Umwandlung

- Daten exportieren, Datenbank mit neuem Zeichensatz erstellen, Daten importieren
  - Kann länger dauern
  - Alle Datentypen werden (soweit möglich) korrekt umgewandelt
  - Kann nahezu bei allen Zeichensatzkombinationen durchgeführt werden
- Datenbankzeichensatz ändern
  `ALTER DATABASE [<db_name>] CHARACTER SET <new cs>`
  - Es wird nur die „Kennung“ geändert, nicht die Daten
  - Metalink Einträge beachten, wenn CLOB Spalten verwendet werden
  - Nur bei bestimmten Kombinationen möglich (z.B. We8iso8859p1 nach we8mswin1252 ist möglich, nicht aber we8iso8859p1 nach we8iso8859p15) [oder Oracle Support in Anspruch nehmen...]

Character set scanner

- Oracle Tool für Konfliktprüfung vor einer Wandlung
- Script `rdbms/admin/csminst.sql` muss eingespielt werden
- Aufruf:
  `csscan system/manager full=y tochar=we8iso8859p15`
- Es wird ein Protokoll erstellt (scan.txt, scan.err, scan.out). Auszug:

```
Table : WAEHRUNG
Column: WAE_SYMBOL
Type  : VARCHAR2(1)
Number of Exceptions         : 1
Max Post Conversion Data Size: 1
ROWID              Exception Type      Size Cell Data(first 30 bytes)
AAAEzpAADAAAAATAAA lossy conversion         €
```

- Korrektur von Daten muss u.U. Separat nach der Migration erfolgen

Wahl des Zeichensatzes

- Alle abzubildenden Zeichen sollten unterstützt werden
- Konvertierung erfordert Rechenleistung
  - Nach Möglichkeit den Zeichensatz verwenden, mit dem die Clients arbeiten, die den größten Traffic-Anteil ausmachen
- UTF16 kann bei Oracle 8i nicht verwendet werden
- UTF16 (UCS2) kann nicht als „database character set“ verwendet werden, nur als „national character set“ (nchar, nvarchar, nclob)
  - UTF16 braucht doppelt so viel Platz wie ISO8859P1
- Metalink konsultieren ob Besonderheiten zu berücksichtigen sind

Oracle und UTF-8 Kodierung

- UTF-8 Kodierung verwendet variable Längen, rückwärtskompatibel mit US7ASCII
- In Oracle 8i ist die Spezifikation nach Unicode 2.1 implementiert, bei den definierten Zeichen werden bis zu 3 Byte pro Zeichen benötigt (Zeichensatzname UTF8)
- In Oracle 9i ist die Spezfikation nach Unicode 3.1 implementiert, bei den definierten Zeichen werden bis zu 4 Byte pro Zeichen benötigt (Zeichensatzname AL32UTF8)
- Deklaration varchar2(xyz) beziehen sich auf Byte, nicht auf Zeichen, d.h. Bei einer Umstellung auf UTF8 muß das Datenbankschema angepasst werden – z.B. Das €-Zeichen benötigt 3 Byte

Resumee

- Für Anwendungsszenarien, wo mehrere unterschiedliche Clients oder mehrere unterschiedliche Datenbanken mit unterschiedlichen Zeichensätzen verwendet werden, ist eine Migration angebracht.
  - Ohne Migration steigt die Anzahl der Workarounds, das kompliziert die Migration
  - Konfiguration von Clients, Schnittstellen usw. muß gleichzeitig mit der DB-Zeichensatzumstellung erfolgen, dies bedeutet einen größeren Aufwand
  - Sensibilisierung der Entscheider notwenig („Es funktioniert ja alles“)

Links

- [1] Metalink, Note 144192.1
- [2] Metalink, Note 68790.1
- [3] Metalink, Note 158577.1
- [4] Metalink, Note 137127.1
- [5] Metalink, Note 119164.1
- [6] Metalink, Note 66320.1
- [7] http://otn.oracle.com/tech/globalization/pdf/Unicode.PDF
- [7] http://www.microsoft.com/globaldev/reference/cphome.mspx
- [8] http://www.cs.tut.fi/%7Ejkorpela/chars/index.html
- [9] http://www.cl.cam.ac.uk/~mgk25/unicode.html#utf-8

Fragen / Diskussion