informacijski inženjering

Razvoj aplikacija na kodnoj stranici UTF8

Stanko Plivelić, dipl.ing.

30.09.2007.

12.konferencija HrOUG, Rovinj, listopad 2007.

O čemu će biti riječi...

- Unicode i UTF8
- podešenja baze
- podešenja klijenta
- iskustva u radu sa UTF8 bazom i Oracle Forms, Reports, SQL Loader-om, exp/imp...
- naglasak na klijentsko-serverskom modu rada

Pojmovi...

- character set
  - skup znakova koji uključuje mapiranje između tih znakova i njima pripadnih numeričkih vrijednosti
  - primjeri: ASCII, Unicode...
- character set encoding
  - kako numeričke vrijednosti pohraniti u računalu
  - primjer: UTF-8
- code page (kodna stranica)
  - tablica za mapiranje znakova u numeričke vrijednosti
  - primjeri: ANSI 1250, OEM 852

...i još malo pojmova

- font
  - skup grafema kojim se znakovi pretvaraju u otisak na ekranu/papiru
  - primjer: Arial – implementira ANSI 1250, 1251, 1252, 1253, 1254, 1257
  - ne mora postojati grafička prezentacija za svaki znak

Malo povijesti...

- ASCII
  - 7 bita (128 znakova)
  - slova samo engleske abecede
- 8-bitne kodne stranice (character set-ovi)
  - ANSI 1250
  - OEM 852
- Unicode
  - trenutno oko 100-ak tisuća znakova, 30-ak alfabeta

Unicode

- svaki znak prikazuje se tzv. code point-om
  - primjer: code point za A je U+0041
- UTF (Unicode transformation format)
  - UTF-8
  - UTF-16
    - 2 bajta (ili 4)
    - UTF-16LE i UTF-16BE
    - UCS-2 – samo 2 bajta
  - UTF-32/UCS-4
    - 4 bajta

UTF-8

- varijabilna duljina prikaza 1 znaka (od 1 do 4 bajta)
  - 1 bajt – ASCII znakovi
  - 2 bajta – npr. hrvatski dijakritički znakovi
  - 3 bajta – npr. ćirilični znakovi
  - 4 bajta – linear B, feničko pismo itd.
- svi ASCII tekstovi/datoteke su ujedno i UTF-8 tekstovi/datoteke

BOM

- kako označiti da je neka datoteka u UTF-8 formatu?
- byte order mark (U+FEFF) na početku datoteke
- UTF-16
  - BE (FE FF) ili LE (FF FE)
- UTF-8
  - ne označava poredak bajtova
  - EF BB BF
  - editori koji ne prepoznaju UTF-8 prikazuje BOM kao znakove ð»¿

Oracle i Unicode

- Oracle podržava UTF-8 od verzije 8i
  - character set-ovi UTF8(od 8i) i AL32UTF8 (od 9i)
- AL16UTF16
  - UTF-16 (od 9i)

Character set baze	Verzija baze	Unicode standard
UTF8	8.0-10g	2.1 (8.0-8.1.6) 3.0 (8.1.7-10g)
AL32UTF8	9.0-10g	3.0 (9.0) 3.1 (9.2) 3.2 (10.1) 4.01(10.2)

Character set baze

- određuje koji sve znakovi mogu biti spremljeni u bazu
- parametar baze NLS_CHARACTERSET
- jednobajtni: US7ASCII, EE8MSWIN1250 (hrvatski znakovi), CL8MSWIN1251 (ćirilica), WE8MSWIN1252, EE8ISO8859P2 itd.
- višebajtni: UTF8 (do maks. 3 bajta po znaku), AL32UTF8 (do 4 bajta)
- nema veze sa kodnom stranicom (npr. ACP u windows-ima) OS-a koji je na serveru

Promjena character set-a

- stari character set treba biti podskup novog – inače može doći do gubitka podataka
- ? - replacement character
- csscan + alter database character set
- duljina imena objekata (30) i usera (8) je u bajtima

NLS_NCHAR_CHARACTERSET

- dodatni character set baze
- za varijable i kolone tipa NCHAR, NVARCHAR2 i NCLOB
- Forms 6i/9i ne podržavaju ove tipove

NLS_LENGTH_SEMANTICS

- što znači npr. varchar2(10), odnosi li se 10 na znakove ili na bajte?
- od 9i moguće je definirati kolone kao:
  - kol1 varchar2(10 byte)
  - kol2 varchar2(10 char)
- ako nije eksplicitno navedeno, gleda se vrijednost NLS_LENGTH_SEMANTICS
- na nivou instance baze ili na nivou sesije
- default-na vrijednost je byte
- ograničenje na 4000 bajta

NLS_LANG (1)

- <jezik>_<teritorij>.<klijentski characterset>
- ako nije definiran <klijentski characterset>, onda ga određuje <jezik>
  - npr. za <jezik> CROATIAN, EE8MSWIN1250
- windowsi:
  - u registry-u pod HKEY_LOCAL_MACHINE\SOFTWARE\ORACLE
  - kao sistemska varijabla sa set NLS_LANG=CROATIAN_CROATIA.EE8MSWIN1250
- ako nije definiran NLS_LANG, <klijentski characterset> je US7ASCII

NLS_LANG (2)

- treba odgovarati character set-u/kodnoj stranici operacijskog sustava koji se vrti na klijentu
- u windows-ima to je ACP (ANSI code page)
  - za "naše" windows-e ACP je 1250 pa klijentski character set treba biti EE8MSWIN1250
  - za DOS-OEM verziju sqlplus-a treba staviti EE8PC852
- NLS_LANG treba podesiti prije pokretanja klijentske aplikacije koja ga koristi
- klijentski character set ne može se promijeniti sa alter session

NLS_LANG (3)

- ako je klijentski character set jednak baznom, ne dolazi do konverzije
- konverzija se obavlja na klijentu (osim kad klijent ne poznaje character set baze)
- ako u klijentskom character set-u ne postoji znak koji treba prikazati, prikazuje se znak ?
  - npr. treba prikazati ćirilični znak, a klijentski character set je EE8MSWIN1250
- u nekim aplikacijama (npr. TOAD) potrebno je podesiti i script za font

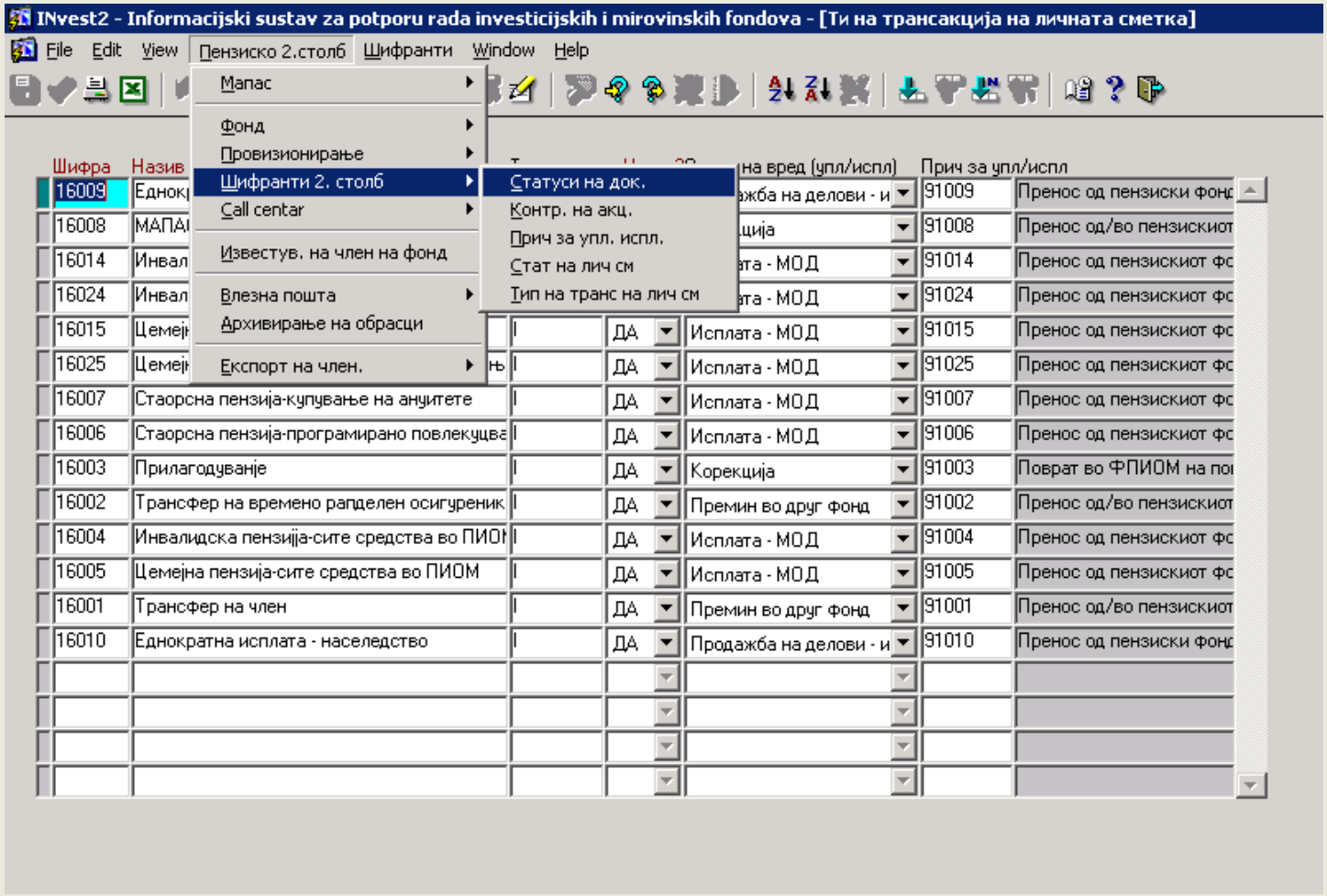
Forms/Reports i UTF8 (1)

- postavljanje klijentskog character set-a na UTF8
  - ako u formi/meniju/izvještaju postoje dijakritički znakovi – može doći do rušenja forme/izvještaja
- rad sa UTF8 bazom
  - punjenje item-a širine 1 na formi znakom koji zahtijeva više od jednog bajta pozivanjem procedure sa out parametrom
    - rješenje: ne koristiti proceduru sa out parametrom
  - out parametar u baznoj proceduri tipa record – javlja se greška 'signature of package has been changed'
    - rješenje: jedna verzija za UTF8 bazu, druga za jednobajtne baze

Forms/Reports i UTF8 (2)

- Forms 9i
  - property item-a Data Length Semantics
- razvoj aplikacija sa različitim character set-ovima/kodnim stranicama na istom računalu
  - promjena kodne stranice zahtijeva restart računala
  - preporuka: dodatno računalo (ili virtual machine)

Primjer forme na ćirilici



SQL Loader i UTF-8

- kakav se character set/encoding koristi u datoteci koja se učitava?
- koriste se CTL parametri CHARACTERSET i BYTEORDER (ako nisu definirani, onda je bitan klijentski character set)
- klijent verzije 9 zna čitati datoteke u UTF-16 formatu, kao i prepoznavati BOM (klijent verzije 8 to ne zna)

Exp/imp i UTF-8

- cilj je smanjiti broj konverzija iz jednog character set-a u drugi
- export i import je najbolje raditi sa klijentskim character set-om postavljenim na istu vrijednost kao što je i character set izvorišne baze
- nije bitna kodna stranica OS-a računala na kojem se radi export/import
- klijenti verzije 8 i stariji kod exporta/importa ne rekonstruiraju ispravno strukturu baze
  - kolone tipa varchar2 proširuju s faktorom 3

NLS u web formama (apleti)

- klijent – iAS - baza
- klijent: npr.Windows(ANSI code point)+browser+JInitiator(ANSI->UTF-16)+Forms aplet(UTF-16->UTF-8)
- iAS(UTF-8->character set definiran u NLS_LANG-u Forms-a)
- baza(eventualna konverzija u OCI sloju)
- preporuka: staviti NLS_LANG za Forms-e jednak character set-u baze

Zaključak

- odabir character set-a ovisi o skupu znakova koje treba podržati
- klijenti verzije 8i i stariji imaju problema u radu sa character set-om AL32UTF8
- uz oprez pri podešavanju NLS_LANG i NLS_LENGTH_SEMANTICS nema većih zapreka za korištenje character set-a UTF8
- višebajtni character set traži veće smještajne (storage) kapacitete

Hvala na pozornosti... :)

... pitanja?