Fokusiranje na kvalitetu u skladištu podataka

Darko Homar
DEKOD telekom d.o.o.
darko.homar@dekod.hr

HrOUG 2007

Cilj

- pregled raznih aspekata vezanih uz kvalitetu podataka
- zašto je kvaliteta podataka važna?
- uzroci loše kvalitete podataka
- mjerenje kvalitete podataka
- ugradnja u proces učitavanja
- prateća organizacija
- dokumentiranje

Uvod

- zadatak skladišta podataka
  - ujedinjavanje podataka iz različitih izvora
  - vremenska dimenzija podataka
- faktori uspjeha
  - relevantnost podataka
  - kvaliteta podataka
  - dostupnost – podaci moraju biti učitani na vrijeme

Uvod

- poteškoće izgradnje skladišta podataka
  - obuhvat cijelog poslovanja poduzeća
  - velik broj poslovnih pravila
  - velik broj atributa i definicija
  - velika količina podataka
  - kratki rokovi implementacije
  - kratko vrijeme učitavanja podataka u skladištu

Uvod

- uspostavljanje skladišta podataka
  - formiran tim
  - infrastruktura: serveri, storage, baza, DBA podrška
  - odabrani alati
  - poslovna analiza i ciljevi (detalji?)
  - dizajn skladišta podataka
  - definirani izvori podataka
  - ETL procedure
  - inicijalno punjenje podataka
  - automatizacija dnevnog ETL-a

Je li posao gotov?

- ogroman posao uspostave skladišta, pritisak na prvi rezultat
- korisnici trebaju iteracije da bi definirali željeno
- korisnici ne poznaju procese i podatke!
- sukob korisnika: oni koji definiraju i oni koji koriste podatke
- posao i podaci se neprekidno mijenjaju

Metapodaci – definicija podataka

- konkretizirajmo problem!
- jednoznačnost, jasnoća definicije
- terminologija unutar kompanije, vanjska terminologija
- primjer: zadnje oročenje depozita - issue date, open date

Sadržaj atributa

- tip podatka
- range
- lista vrijednosti
- poseban format, npr. telefonski broj
- null – not null

Sadržaj atributa - primjer

- podatak o spolu – M ili F
  - transformacije:
    - m,M,muško, male => M
    - ž, Ž, F, female, žensko => F
  - pravne osobe => N/A
  - što se može naći u izvoru podataka?
    - fizička osoba bez informacije o spolu
    - pravna osoba s informacijom o spolu
    - fizička osoba s nerazumljivom oznakom, npr. “Z”

Sadržaj atributa - primjer

- datumsko polje
  - neispravan datum, npr. 29.2.2003.
  - nepostojeći datum, npr “00000000”, ovo može biti i oznaka za NULL datum
  - format varira 15.05.2003, 03/05/15...

Poslovna pravila

- određuju međuovisnost atributa u tablici
- određuju međuovistnost atributa u različitim tablicama
  - npr: ako je klijent pravna osoba, spol je “N/A”
  - npr: odobreni, a neisplaćeni kredit zabilježen je na kontima vanbilance
  - npr: kredit s valutnom klauzulom ne može biti vezan uz domicilnu valutu

Pravila koja izviru iz strukture skladišta podataka

- objekti u DW prate povijesne promjene
- na primjer, polja u tablici klijenata:
  - UNIFIED_KEY – unificirani ključ (šifra) klijenta
  - NAME – naziv klijenta
  - DAT_FROM – datum od kada vrijedi ovaj podatak
  - DAT_TO – datum do kada vrijedi podatak

Primjer

UNIFIED_KEY	NAME	DAT_FROM	DAT_TO
12345	Izgubljeno d.o.o.	01.05.2006	01.12.2006
12345	Nađeno d.o.o.	01.12.2006	31.12.9999

digresija: ispravak u odnosu na izmjenu

Preklapanje intervala

UNIFIED_KEY	NAME	DAT_FROM	DAT_TO
12345	Izgubljeno d.o.o.	01.05.2006	01.12.2006
12345	Nađeno d.o.o.	30.11.2006	31.12.9999

Rupe u intervalima

UNIFIED_KEY	NAME	DAT_FROM	DAT_TO
12345	Izgubljeno d.o.o.	01.05.2006	01.12.2006
12345	Nađeno d.o.o.	01.01.2007	31.12.9999

UNIFIED_KEY	NAME	DAT_FROM	DAT_TO
12345	Izgubljeno d.o.o.	01.05.2006	01.12.2006
12345	NULL	01.12.2007	01.01.2007
12345	Nađeno d.o.o.	01.01.2007	31.12.9999

Referencijalni integritet

UGOVOR

UNIFIED_KEY	KEY_CTR	KEY_PDT	IR_TYPE	MAT_DATE	DAT_FROM	DAT_TO
22222	12345	P0101	fixed	01.01.2007	01.04.2006	31.12.9999
22222	12345	P0101	fixed	01.01.2008	15.12.2006	31.12.9999

KLIJENT

UNIFIED_KEY	NAME	DAT_FROM	DAT_TO
12345	Izgubljeno d.o.o.	01.05.2006	01.12.2006
12345	Nađeno d.o.o.	01.12.2006	31.12.9999

Uzroci loše kvalitete podataka

- inherentni
  - transakcijski sustavi ne pamte prošlost
    - npr. izmjena cjenika
  - poteškoće u punjenju FDW-a
    - posebno inicijalno punjenje
    - podaci se nadopunjuju, što unosi nesigurnost, tj. smanjuje kvalitetu podataka

Uzroci loše kvalitete podataka

- nedostatak kontrole na izvoru
  - nemogućnost generalnog predviđanja kvalitete podataka u izvoru
  - najsigurnija pretpostavka: ne vjerovati u definiciju izvora!
    - datumi: 31.4.2007
    - 43,5612 kn
    - m, ž, d (dijete)
- poslovna pravila – još gora situacija
  - npr. kriva knjiženja, npr. pripajanje poduzeća

Uzroci loše kvalitete podataka

- Propusti u standardizaciji skladišta podataka
  - neugodna tema za nas, “graditelje skladišta”
  - primjer: punjenje klijenta iz 4 izvora
  - stotine mapiranja, tisuće atributa
  - iskustvo i promjene u timu
  - nesavršeni alati
  - nedostatak koncentracije i kontrole

Uzroci loše kvalitete podataka

- bugovi
  - bugovi ne postoje u planovima
  - mogu snažno utjecati na softverska rješenja
  - mogu utjecati na podatke u skladištu
    - pada povjerenje u sustav
    - rješenja se moraju čekati
    - zaobilazna rješenja mogu biti radikalna i skupa

Uzroci loše kvalitete podataka

- nepoznavanje izvora - prilagođeni Murphy:
  - ne postoji dobro dokumentiran informatički sustav
  - ako postoji dokumentacija, onda ne opisuje postojeći sustav, nego neki zamišljeni
  - ako postoji dokumentacija postojećeg stanja, onda je tolika da se ne može savladati u vremenu dostupnom smrtnicima
  - postoje ljudi koji poznaju rad sustava
  - ti ljudi ne mogu sustavno i cjelovito opisati rad sustava
  - ali mogu opisati najbizarnije moguće slučajeve, koji se možda i ne događaju
  - ima i preciznih: oni odgovaraju s “DA” i “NE
- nerazumijevanje = krivo mapirani podaci
- najbolji razlog za traženje iskusnih konzultanata

Uzroci loše kvalitete podataka

- kriva ili nepotpuna definicija
  - ponovo, miješanje dva velika pitanja skladišta podataka: DEFINICIJA i KVALITETA
  - velika potrošnja vremena (novaca)
    - pokušaj izrade učitavanja prema lošoj definiciji
    - druga krajnost – natezanje definicije da odgovara podacima
  - nema recepta, osim velikog iskustva i povjerenja u kvalitetu ljudi koji rade na skladištu podataka

Upravljanje kvalitetom podataka

- procjena kvalitete podataka
- dizajn pravila
- transformacija
- praćenje kvalitete podataka

Procjena kvalitete podataka

- inicijalno se uzima uzorak podataka i analizira
- često vodi u iteracije: definicija – analiza
  - prisjećanje na implementacijske odluke u transakcijskim sustavima
  - npr. koja je default vrijednost datuma dospijeća za proizvode poput depozita po viđenju

Faktori kvalitete podataka

- težimo numeričkim mjerama kvalitete podataka
- npr:
  - potpunost (completeness)
  - točnost (exactness)
  - ispravnost (validity)
  - preciznost (precision)
  - konzistentnost (consistency)
  - timing

Faktori kvalitete podataka

- ključno: definicija pravila za mjerenje faktora kvalitete
- Oracle Warehouse Builder - “mehanička” analiza:
  - predloženi tip podataka
  - srednja vrijednost, minimalna, maksimalna...
  - kandidati za liste vrijednosti
  - analiza jedinstvenosti sadržaja atributa
  - analiza popunjenosti atributa...

Dizajn pravila i transformacija

- iteracije analize i definicije daju sliku što se želi imati u skladištu podataka i koji su problemi
- dizajniraju se pravila za čišćenje i implementiraju u ETL
- pravila za kontrolu ispravnosti podataka
  - nepoštivanje pravila=izuzetak
- uočiti: tolerancija određene razine problematičnih podataka (pogotovo povijesno)

Praćenje kvalitete podataka

- izostanak praćenja = erozija kvalitete podataka
- kontinuirano praćenje i bilježenje podataka koji ne odgovaraju definicijama
  - definicija:
    - govori što podaci znače
    - zgodno je da govori i o tome što podaci nisu
- podsustav unutar skladišta podataka
  - zahtijeva znatne računalne i organizacijske resurse
  - ako je dobar, signalizira promjene u kvaliteti
  - troši skupo vrijeme za ETL

Ugradnja DQ u skladište podataka

- u idealnom svijetu:
  - u skladištu nema loših podataka
  - ispravak u ETL transformacijama
  - izuzeci
    - zaustavljanje ETL procesa
    - ispravljanje
    - učitavanje ispravljenih podataka

Ugradnja DQ u skladište podataka

- u realnom svijetu
  - stotine svakodnevnih ETL procedura
  - zaustavljanje procesa je luksuz
  - svemoguće transformacijske procedure ne postoje
- ostaju dva rješenja:
  - podaci se učitavaju u skladište podataka
  - podaci se “preskaču”, ispravljaju i naknadno učitavaju
  - oba rješenja daju privremeno iskrivljenu sliku

Ugradnja DQ u skladište podataka

- gdje ugraditi kontrolu kvalitete podataka?
  - sastavni dio učitavanja podataka
    - kritični problemi – zaustavljanje procesa učitavanja
    - ostalo – dojavljivanje, logiranje, eventualno preskakanje
  - zaseban proces provjere kvalitete podataka
    - tamo gdje je prevelik utjecaj na performanse
    - tamo gdje je procjena da neće biti kritičnih problema
    - tamo gdje su procesi učitavanja nezavisni (različiti izvori), a podaci su povezani

Organizacija za podršku DQ

- uzroci loše kvalitete podataka uglavnom nisu u timu za skladište podataka
- uzroci se kriju u bilo kojem dijelu organizacije
- skladište podataka – izvrsno mjesto za detekciju problema s podacima
- skladište podataka u principu ne smije ispravljati podatke

Organizacija za podršku DQ

- skladište podataka ne odgovara za nevaljale podatke iz izvora
- transakcijski sustavi imaju dovoljno dobre podatke za svoje potrebe
- korisnici skladišta nemaju dobar uvid ni ovlasti za uklanjanje problema
- posljedica: ništa se ne poduzima za poboljšanje kvalitete podataka

Organizacija za podršku DQ

- zamislimo organizacijsku funkciju sa svrhom poboljšanja kvalitete podataka – jedinstveni centar kontrole
- zadaci
  - zaprimanje problema s kvalitetom podataka
  - pronalaženje izvora problema
  - praćenje i osiguravanje ispravljanja podataka
  - praćenje i osiguravanje uklanjanja izvora problema
  - vraćanje ispravljene informacije u skladište podataka

Metapodaci - dokumentiranje

- srodnost sadržaja i definicije podataka
- napomene o kvaliteti trebaju biti zapisane zajedno s definicijom podataka
  - mogu se navesti izuzeci, defaultne vrijednosti
  - važna transformacijska pravila
  - procjena kvalitete (koliko podataka nije u skladu s definicijom)

Zaključak

- pregled problema kvalitete podataka
- uzroci loše kvalitete
- pokušaj ocrtavanja cijelog procesa poboljšanja kvalitete
- ugradnja DQ u infrastrukturu skladišta podataka
- organizacija za poboljšanje kvalitete podataka
- metapodaci - dokumentiranje

Fokusiranje na kvalitetu u skladištu podataka

Darko Homar
DEKOD telekom d.o.o.
darko.homar@dekod.hr

HrOUG 2007