

Održavanje podataka u Data Warehouse sistemima

PREUZIMANJE, TRANSFORMACIJA I
PUNJENJE PODATAKA U DW BP

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Procesi održavanja podataka u DW

Tipične aktivnosti projektovanja DW sistema

- analiza i specifikacija korisničkih zahteva
 - specifikacija poslovnog modela
- projektovanje šeme DW BP
 - projektovanje konceptualne šeme DW BP
 - specifikacija logičkog modela
 - projektovanje implementacione šeme DW BP
 - specifikacija dimenzionog modela
 - projektovanje fizičke organizacije šeme DW BP
 - specifikacija fizičkog modela
- projektovanje arhitekture DW sistema
- **projektovanje ECTL softverske podrške**
- projektovanje softverske podrške za izveštavanje i analizu podataka

Procesi održavanja podataka u DW

Kreiranje i osvežavanje DW

- ETL (ECTL) proces

Softverska podrška za zadatke

- **Extraction (ekstrakcija)**
 - selektovanje (izdvajanje, preuzimanje) podataka iz različitih izvora
- **Cleaning & Transformation (transformacija)**
 - validacija, prečišćavanje, integracija i vremensko označavanje podataka
- **Loading (punjenje)**
 - punjenje DW baze podataka
 - inicijalno
 - regularno osvežavanje

Procesi održavanja podataka u DW

Kreiranje i osvežavanje DW

- ETL (ECTL) proces

Postoje različita softverska rešenja i pristupi za podršku ovih zadataka

- upotreba gotovih softverskih paketa
 - vendor ETL tools
- upotreba sopstvenih softverskih rešenja
 - in-house ETL tools

Procesi održavanja podataka u DW

Moguće tehnološke osnove za ETL (1/5)

- direktna upotreba 3GL programskih jezika
- upotreba utility softverskih alata
- upotreba naprednih mogućnosti jezika SQL
- upotreba posrednika - Gateway interfejsa

Procesi održavanja podataka u DW

Moguće tehnološke osnove za ETL (2/5)

- **direktna upotreba 3GL programskih jezika**
 - npr. razvoj i upotreba C, C++, Java, PL/SQL, T-SQL, Cobol programa za direktni pristup podacima
 - memorisanim u BP ili datotekama

Procesi održavanja podataka u DW

Moguće tehnološke osnove za ETL (3/5)

- **upotreba utility softverskih alata**
 - programa tipa Export, Import ili Load
 - mogućnost prenosa opisa šeme BP i različito formatiranih podataka, npr. binarno, tekstualno, XML, JSON
 - binarni oblik je zavisn od interne organizacije izabranog SUBP
 - mehanizama SUBP za replikaciju ili asinhroni prenos poruka
 - mehanizama za fizičko kopiranje fajlova sa podacima
 - Primer Oracle: tehnika "transportable tablespace" i FTP
 - SQL mehanizama za kreiranje i pristup eksternim tabelama
 - eksterna tabela: tabela sa podacima koji se fizički nalaze u eksternoj datoteci, umesto u BP
 - jedina dopuštena operacija je selektovanje podataka
 - Primer Oracle:
 - CREATE TABLE ... (...) ORGANIZATION EXTERNAL (...)

Procesi održavanja podataka u DW

Moguće tehnološke osnove za ETL (4/5)

- **upotreba naprednih mogućnosti jezika SQL**
 - kombinacija kreiranja tabele i istovremenog selektovanja podataka iz drugih tabela (CTAS)
 - CREATE TABLE ... (...) AS SELECT ...
 - masovno upisivanje torki u tabelu selektovanjem podataka iz drugih tabela (ITAS)
 - INSERT INTO TABLE ... AS SELECT ...
 - masovno upisivanje torki u više tabela selektovanjem podataka iz drugih tabela
 - multitable insert i pivoting insert
 - INSERT [ALL | FIRST] [[WHEN ...] INTO ...]... SELECT...
 - masovno modifikovanje torki na osnovu selekcije podataka
 - uslovno upisivanje ili modifikacija torki na osnovu selekcije
 - MERGE INTO...WHEN MATCHED... WHEN NOT MATCHED...

Procesi održavanja podataka u DW

Moguće tehnološke osnove za ETL (5/5)

- **upotreba posrednika - Gateway interfejsa**
 - Gateway – API za prosleđivanje SQL naredbi iz klijentskog programa prema SUBP
 - Tehnologije pristupa podacima
 - Open Database Connectivity (ODBC)
 - Java Database Connectivity (JDBC)
 - Object Linking and Embedding for Databases (OLE DB)
 - ActiveX Data Objects .NET (ADO .NET)
 - kombinovane s tehnologijama objektno-relacionog preslikavanja, eng. Object Relational Mapping (ORM)
 - Hibernate
 - Entity Framework

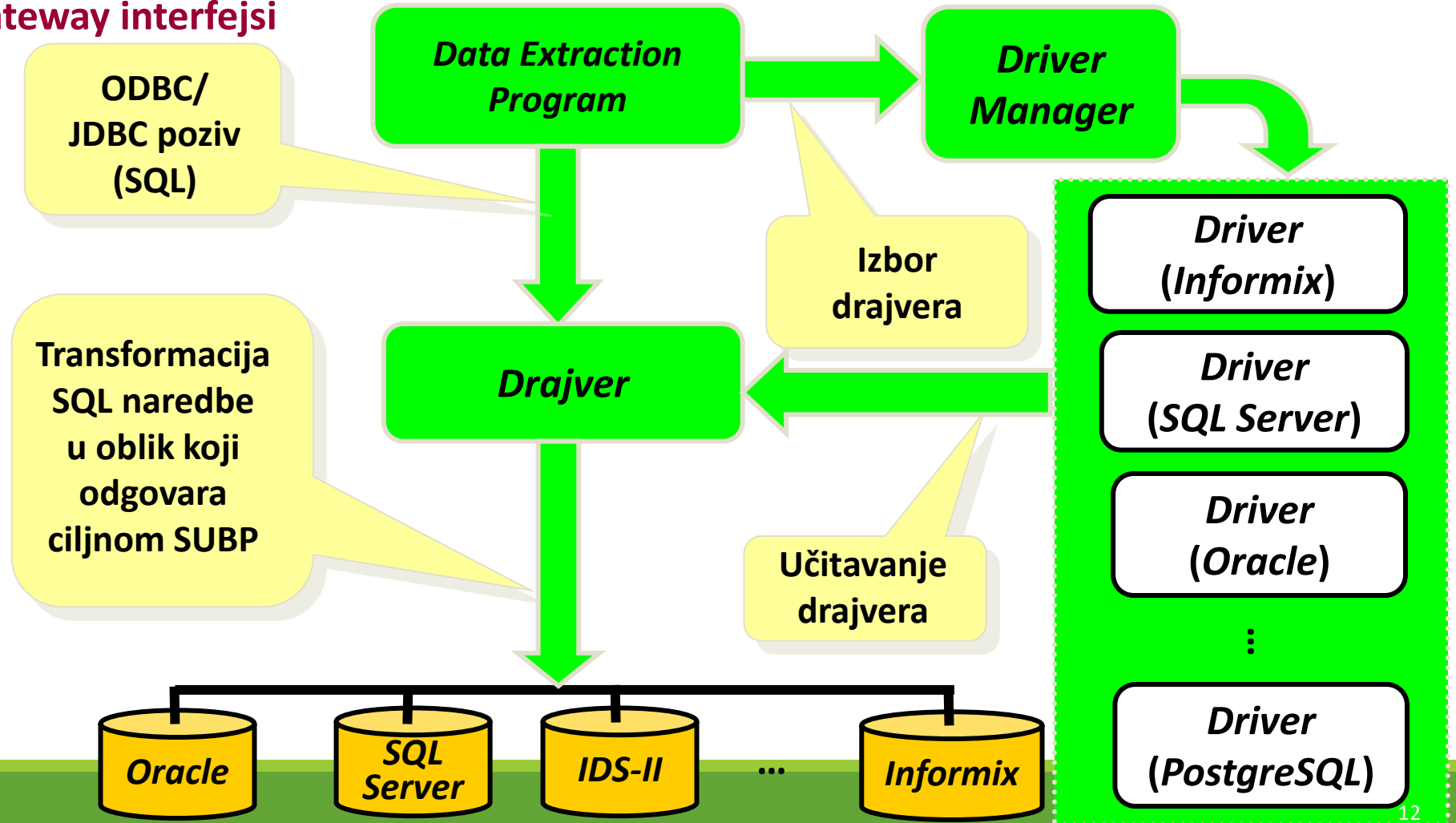
Procesi održavanja podataka u DW

Posrednici - Gateway interfejsi

- omogućavanje konekcije programa na gotovo bilo koji SUBP
- uvode novi nivo indirekcije u prosleđivanju SQL naredbi
 - podržan drajverom za izabrani, ciljani SUBP
 - drajver prevodi standardnu SQL naredbu u oblik specifičan za izabrani SUBP
- program (i izvorni i izvršni) postaje nezavisan od pozadinskog SUBP
- pogodno rešenje za slučajeve kada
 - je potrebna online ekstrakcija iz baze podataka
 - se ekstrahuje manja količina podataka
 - problem performansi u slučaju ekstrakcije velikih količina

Procesi održavanja podataka u DW

Posrednici - Gateway interfejsi



Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Ekstrakcija podataka

Selekcija i preuzimanje izvornih podataka

Mogući izvori podataka

- **produkcioni podaci**
 - OLTP sistemi, podržani putem SUBP, NoSQL, ili sistema datoteka
- **arhivski podaci**
 - arhivske kopije podataka
 - najčešće za obuhvat dugačkih istorijskih perioda vremena
 - pogodne za inicijalno punjenje DW BP
- **interni izvori podataka**
 - "privatni" podaci u fajlovima različitih formata, npr. CSV, XML, JSON, .xls, .doc, slabo ili dobro strukturirani
- **eksterni izvori podataka**
 - pribavljeni izvan organizacionog sistema

Ekstrakcija podataka

Moguće vrste ekstrakcije podataka

- sa stanovišta preuzetih podataka
 - **potpuna**
 - uvek se preuzimaju kompletni podaci izvora
 - **inkrementalna**
 - preuzimaju se samo novopridodati podaci izvora, u odnosu na vremenski trenutak završetka prethodne ekstrakcije

Ekstrakcija podataka

Moguće vrste ekstrakcije podataka

- sa stanovišta pristupa izvoru podataka
 - **u radnom režimu (online)**
 - pristup izvoru u operativnoj upotrebi
 - npr. pristup serveru BP koji je u radnom režimu
 - **van radnog režima (offline)**
 - pristup izvoru van operativne upotrebe
 - datotekama podataka
 - log, arhivskim, trace, ili dump datotekama

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Transformacija podataka

Cleaning & Transformation (transformacija)

- validacija, pročišćavanje, integracija i vremensko označavanje podataka
- često najkompleksniji i najzahtevniji deo ETL procesa
- realizuje se u okviru Data Staging Area (DSA) područja
- ima ključni uticaj na obezbeđenje kvaliteta podataka u DW sistemu
 - zahtev za definisanje **politike i sistema obezbeđenja kvaliteta** u izgradnji i održavanju DW sistema
 - moraju biti što preciznije definisati kriterijumi zahtevanog kvaliteta podataka u DW sistemu
 - uvođenje nove uloge u timu: **Data Quality Manager**

Transformacija podataka

Pozicioniranje DSA područja

- DSA zajedno sa OLTP serverom
 - **on-site staging model**
- DSA na posebnom serveru
 - **standalone remote staging model**
- DSA zajedno sa DW serverom
 - **remote staging model**

Pozicioniranje postupaka transformacije

- paralelno sa ekstrakcijom podataka
- nakon ekstrakcije i pre punjenja podataka u DW BP
- paralelno sa punjenjem podataka u DW BP
- kombinacija prethodno navedenih mogućnosti

Transformacija podataka

Prečišćavanje podataka

- usled visoke verovatnoće pojave neusaglašenosti, anomalija ili grešaka nad podacima iz različitih izvora

Zadaci prečišćavanja i transformacije podataka

- otkrivanje i otklanjanje grešaka i neusaglašenosti
- transformacija podataka u oblik pogodan za punjenje u DW
 - opremanje podataka vremenskom dimenzijom i svođenje na istu vremensku osu
 - transformisanje podataka iz jednog u drugi format
 - spajanje različito identifikovanih torki u jednu, ili razdvajanje jedne torke na više različitih, s različitom identifikacijom
- integracija podataka - spajanje isto identifikovanih torki u jednu

Transformacija podataka

Mogući tipovi neusaglašenosti ili grešaka (1/3)

- **nepostojanje ili nepoštovanje ograničenja ključa**
 - nekonzistentna identifikacija torki u relacijama
 - pojava različitih torki sa istom vrednošću ključa
 - pojava torki sa različitim vrednostima ključa koje opisuju isti realni objekat
- **nepostojanje ili nepoštovanje drugih ograničenja**
 - narušavanje ograničenja vrednosti atributa, ograničenja torke, ograničenja referencijalnog integriteta, itd.
- **pojava atributa sinonima i atributa homonima**
 - neusaglašenost naziva i semantike, pridružene atributima
 - jedan realno isti atribut dekomponovan u više različitih, ili
 - više realno različitih atributa objedinjenih u jedan
 - tipični slučajevi: načini formatiranja poštanskih adresa ili imena ljudi

Transformacija podataka

Mogući tipovi neusaglašenosti ili grešaka (2/3)

- **neusaglašenost ograničenja domena istih atributa**
 - neusaglašenost tipova podataka, dužina, logičkih uslova, kodnih rasporeda i formata zapisa vrednosti
- **neusaglašenost istih vrednosti atributa**
 - neusaglašenost korišćenih znakova
 - velika/mala slova, {š, č, ć, đ, ž}/{s, c, c, dj, z}, latinica/ћирилица
 - neusaglašeno kodiranje vrednosti
 - različito kodirana ista vrednost, npr. BGD/BEG – Beograd
 - isto kodirane različite vrednosti, npr. SK – Sisak/Skoplje
 - neusaglašene jedinice mere vrednosti
 - ista vrednost data u različitoj jedinici mere, npr. 1 EUR / 120 RSD?
 - slučajne greške pri unosu podataka (tipfeleri)

Transformacija podataka

Mogući tipovi neusaglašenosti ili grešaka (3/3)

- **neusaglašenost vrednosti ključa u skupu realnih objekata**
 - nekonzistentna identifikacija istog objekta u sistemu – nekonzistentno definisane vrednosti ključa unutar klase objekata
- **izostavljeni podaci**
 - nedostajuće celokupne torke u izvorima podataka ili
 - nedostajuće (null) vrednosti atributa u izvorima podataka
- **ekstremni podaci**
 - Vrednosti atributa koje su izvan skupa dozvoljenih vrednosti
 - Neobične ili neočekivane vrednosti atributa
 - Još uvek unutar skupa dozvoljenih vrednosti, ali neočekivano ekstremne

Transformacija podataka

Mogući pristupi u razrešavanju neusaglašenosti ili grešaka

- **tolerisati greške i neusaglašenosti**
 - preuzimanje "prljavih" podataka u DW - ne preporučuje se
- **ignorirati greške i neusaglašenosti**
 - "prljavi" podaci se ignorišu – ne preuzimaju u DW
 - ponekad može biti tolerantan pristup
- **suštinski i formalno razrešiti greške i neusaglašenosti u ETL postupku**
 - realno najteži pristup, ali suštinski vodi ka obezbeđenju zahtevanog kvaliteta podataka u DW sistemu
 - zahteva jasnu uspostavu odgovornosti nad podacima i postupcima prečišćavanja i transformacije
 - zahteva intenzivnu saradnju i članova tima i korisnika

Transformacija podataka

Transformacija podataka za punjenje u DW (1/3)

- **opremanje podataka vremenskom dimenzijom i svođenje na istu vremensku osu**
 - OLTP sistemi i drugi izvori podataka retko sadrže vremensku dimenziju koja je "poravnata", dobro definisana i obuhvata istorijski dugačak period vremena
 - svodi se na zapis datuma i vremena nastanka promene podataka u sistemu
 - uobičajeno je da se vremenska dimenzija sa hijerarhijama u DW sistemu pripremi unapred, pre inicijalnog punjenja
 - vremenskom dimenzijom mogu se opremiti i činjenice i dimenzije
 - dozvoljeno je evidentiranje
 - diskretnih vremenskih trenutaka i
 - vremenskih intervala

Transformacija podataka

Transformacija podataka za punjenje u DW (2/3)

- **spajanje različito identifikovanih torki u jednu**
 - posledica opredeljenja za grublju granularnost po određenoj dimenziji, u odnosu na analizirani izvor podataka ili
 - posledica potrebe agregiranja podataka o istom realnom objektu, u datom vremenskom intervalu
 - npr. opredeljenje da se agregiraju podaci o izvršenim transakcijama i storno transakcijama u zadanom vremenskom intervalu u jednu torku
 - umesto da se svaka torka o svakoj transakciji preuzima posebno
 - posledica razrešavanja neusaglašenosti ili nepostojanja ograničenja ključa

Transformacija podataka

Transformacija podataka za punjenje u DW (3/3)

- **razdvajanje jedne torke na više različitih, s različitom identifikacijom**
 - posledica opredeljenja za finiju granularnost po određenoj dimenziji, u odnosu na analizirani izvor podataka
 - posledica razrešavanja neusaglašenosti ili nepostojanja ograničenja ključa
- **integracija podataka - spajanje isto identifikovanih torke u jednu**
 - posledica objedinjavanja podataka o istom realnom objektu iz različitih izvora podataka

Transformacija podataka

Alati za prečišćavanje i transformaciju podataka

- zasnovani na već pobrojanim tehnološkim osnovama ETL procesa
- mogu predstavljati ugrađene funkcionalnosti u druge alate za ETL proces
 - ukoliko se u postupku ekstrakcije ili punjenja obavljaju i neki zadaci prečišćavanja ili transformacije podataka
- **Data migration tools**
 - omogućavaju primenu jednostavnih operacija za transformaciju podataka, npr.
 - preslikavanje vrednosti kolone Surname u LastName ili
 - konverziju iz jedne u drugu jedinicu mere, npr. inča u centimetre

Transformacija podataka

Alati za prečišćavanje i transformaciju podataka

- **Data scrubbing tools**

- sofisticiraniji alati - u stanju da implementiraju i primenjuju specificirana pravila iz domena primene, npr.
 - da primene validaciju fz ProizID → ProizNaziv za prečišćavanje podataka o proizvodima iz dva različita izvora
 - da konvertuju kodiranu vrednost u originalnu vrednost, npr. kod 381 u 'Srbija' ili
 - da vrše parsiranje i transformaciju podataka, npr. o poštanskim adresama

- **Data auditing tools**

- primenjuju tehnike istraživanja podataka (Data Mining) u pronalaženju netipičnih uzoraka podataka, npr.
 - ekstremno velikih ili malih, ali dozvoljenih vrednosti atributa
 - proizvoda koji gotovo nikada nisu bili prodavani, itd.

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Punjenje podataka

Data Loading

- punjenje DW baze podataka
 - inicijalno
 - regularno osvežavanje

Punjenje podataka

Inicijalno punjenje

- jednokratna procedura preuzimanja istorijskih podataka i početnog formiranja DW BP
 - i dimenzija i činjenica
- obuhvata izrazito veliku količinu podataka
- zahteva potpunu ekstrakciju i posebne procedure prečišćavanja podataka
- zahteva kompleksnu obradu pratećih podataka u pretprocesiranju ili postprocesiranju
- može duže trajati
 - pošto je jednokratno i odvija se u postupku formiranja DW, ne postoji veliki pritisak za kratkim vremenom trajanja

Punjenje podataka

Moguće vrste punjenja podataka

- sa stanovišta preuzetih i transformisanih podataka
 - **potpuno**
 - uvek se pripremaju i pune kompletni podaci u DW
 - vremenski i resursno zahtevan pristup
 - ukoliko je DSA odvojen od DW sistema, to dozvoljava neometan operativni rad korisnika DW sistema
 - **inkrementalno**
 - pripremaju se i pune samo novopridodati podaci u DW, u odnosu na vremenski trenutak završetka prethodnog punjenja
 - vremenski i resursno manje zahtevan pristup
- sa stanovišta pristupa DW sistemu
 - **van radnog režima (offline)**
 - pristup DW sistemu van operativne upotrebe
 - da bi se garantovalo obezbeđenje konzistentne slike podataka

Punjenje podataka

Alati za punjenje podataka

- zasnovani na već pobrojanim tehnološkim osnovama ETL procesa
- insistira se na mogućnosti primene arhitektura i tehnika paralelizacije i vremenskog smicanja (pipelining-a) operacija punjenja podataka
- jer uobičajeno treba obraditi veliku količinu podataka u relativno kratkom vremenu
 - sekvencijalna primena operacija punjenja bi mogla trajati i danima, što je netolerantno dugo vreme
- pogodna primena SUBP koji pružaju mogućnosti paralelizacije SQL i Load operacija nad podacima
 - pSELECT, pCTAS, pITAS, pINSERT, pUPDATE, itd.

Punjenje podataka

Prateći zadaci pri punjenju podataka

- uređivanje (sortiranje) podataka
- agregacija (sumiranje) podataka
- indeksiranje podataka
- kreiranje materijalizovanih pogleda

Administrativni zadaci pri punjenju podataka

- arhiviranje (backup) kreirane ili osvežene DW BP
- punjenje ili osvežavanje meta podataka
- publikovanje napunjenih ili osveženih podataka

Punjenje podataka

Vreme realizacije pratećih zadataka

- generisanje podataka u pretprocesiranju
 - podaci se kompletno pripremaju pre punjenja, u okviru DSA
- generisanje podataka paralelno u toku punjenja
 - podaci se pripremaju u toku punjenja u DW BP
- generisanje podataka u postprocesiranju
 - podaci se pripremaju nakon punjenja u DW BP, na samom DW serveru

Punjenje podataka

Load Window

- minimalno vreme, neophodno za realizaciju svih zadataka punjenja podataka
- nastoji se da bude što kraće
- izbor vremena realizacije pratećih zadataka
 - **generisanje podataka u pretprocesiranju**
 - praktikuje se uvek kada je moguće sve podatke pripremiti unapred u DSA (npr. kod potpunog punjenja)
 - **generisanje podataka paralelno u toku punjenja**
 - npr. generisanje indeksa
 - ne savetuje se, jer znatno produžava load window
 - **generisanje podataka u postprocesiranju**
 - praktikuje se kada je moguće agregirane podatke inkrementalno osvežavati (direktno u DW sistemu) i
 - kada je u pitanju ažuriranje manjeg obima podataka

Punjenje podataka

Load Window

- saveti za skraćivanje load window-a
 - obezbediti očuvanje svih ograničenja podataka u DW sistemu tokom prečišćavanja i transformacije
 - tako što će ETL programi garantovati očuvanje svih deklariranih ograničenja na nivou DW šeme BP
 - isključiti sve mehanizme kontrole ograničenja na nivou DW BP, pri punjenju podataka
 - čime se značajno ubrzavaju naredbe ažuriranja DW BP ili primena Load utility programa
 - izabrati NOLOGGING opciju prilikom punjenja ili osvežavanja DW BP
 - čime se izbegava upisivanje podataka u transakcioni dnevnik i time značajno ubrzavaju naredbe ažuriranja DW BP ili primena Load utility programa

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Osvežavanje podataka

Regularno osvežavanje DW BP

- regularno ponovljiva procedura preuzimanja izmenjenih podataka i osvežavanja DW BP
 - često samo činjenica, ukoliko su dimenzije sporo promenljive
- obuhvata manju količinu podataka
- najčešće zahteva inkrementalnu ekstrakciju i jednostavnije procedure prečišćavanja podataka
- zahteva manje kompleksnu obradu pratećih podataka u pretprocesiranju ili postprocesiranju
- mora relativno kratko trajati
 - npr. realizuje se u toku noći, van radnog vremena
 - često jak pritisak za prilično kratkim vremenom trajanja, kako bi sistem bio 100% raspoloživ u radno vreme korisnika

Osvežavanje podataka

Osnovne tehnike realizacije osvežavanja

- **propagacijom ažuriranja**
 - korišćenjem mehanizama za osvežavanje repliciranih podataka, kao što su trigeri ili materijalizovani pogledi
 - mogu se propagirati operacije, tj. transakcije, ili podaci
- **inkrementalnim punjenjem**
 - preuzimanjem inkrementa podataka iz izvora i punjenjem u DW BP
- izvor može biti OLTP BP, ili DSA područje
- u oba slučaja, nastoji se da se osvežavanje lokalizuje samo na podatke koji su bili predmet izmena
 - kako bi se količina podataka neophodna za obradu pri osvežavanju redukovala na minimum

Osvežavanje podataka

Karakteristike osvežavanja

- load window
- frekvencija i obim podataka za osvežavanje
- pristup detektovanju promena u izvornim podacima
- tehnika i postupak osvežavanja

Osvežavanje podataka

Karakteristike osvežavanja

- **load window**
 - ograničenje maksimalno dozvoljenog vremena trajanja osvežavanja
- **frekvencija i obim podataka za osvežavanje**
 - periodičnost sprovođenja postupka osvežavanja
 - izbor kraćeg perioda znači
 - manju obrađenu količinu podataka pri osvežavanju
 - moguć kraći load window
 - bolju ažurnost podataka u DW BP, ali
 - intenzivnije angažovanje računarskih i mrežnih resursa

Osvežavanje podataka

Karakteristike osvežavanja

- **pristup detektovanju promena u izvoru**
 - šta se smatra promenom i koji su podaci promenljivi
 - na koji način se evidentiraju promene i vremena nastanka promena
- **tehnika i postupak osvežavanja**
 - bitan uticaj na obezbeđenje konzistentne slike podataka u DW
 - kako bi korisnici DW BP uvek videli konzistentu sliku BP

Osvežavanje podataka

Detekcija promena u izvoru podataka = Detekcija inkrementa podataka

- detekcija novih podataka činjenica
 - nove činjenice se, u principu, samo dodaju u DW BP
 - jer obuhvataju prethodno nepokriveni period vremena, a vreme egzistira kao posebna dimenzija
- detekcija novih ili izmena postojećih dimenzionih podataka
 - mogući različiti pristupi u evidentiranju izmena dimenzionih podataka
 - videti temu "Strukture šeme BP DW sistema"

Osvežavanje podataka

Tehnike osvežavanja

- **propagacija ažuriranja**
 - putem trigera baze podataka
 - putem materijalizovanih pogleda i replikacije
 - putem replikacije zasnovane na tokovima podataka
 - data streams tehnika
- **inkrementalno punjenje**
 - zamenom kompletnih podataka
 - poređenjem instanci u odnosu na prethodno osvežavanje
 - sa vremenski obeleženim razlikama
 - upotrebom transakcionog dnevnika

Osvežavanje podataka

Osvežavanje putem trigera baze podataka

- vrste s obzirom na vreme ažuriranja
 - **asinhrono (odloženo)**
 - odloženo, u posebnoj transakciji
 - **sinhrono (trenutno)**
 - trenutno, u istoj transakciji
- vrste s obzirom na način propagacije
 - **tehnika propagacije podataka**
 - propagiraju se novi podaci u DW BP
 - **tehnika propagacije operacija transakcije**
 - propagiraju se operacije transakcija, tako da budu izvršene nad DW BP

Osvežavanje podataka

Osvežavanje putem trigera baze podataka

- **tehnika propagacije podataka**
 - svako ažuriranje izvorne tabele izaziva pokretanje *after row level* trigera
 - koji zapisuje novu sliku podataka u posebnu log tabelu ili fajl
 - posebna procedura osvežavanja propagira izmenjene podatke u DW BP
- **tehnika propagacije operacija transakcije**
 - svako ažuriranje izvorne tabele izaziva pokretanje *after row level* trigera
 - koji propagira sam tekst SQL DML naredbe za izvršavanje nad DW BP

Osvežavanje podataka

Osvežavanje putem materijalizovanih pogleda i replikacije

- upotreba materijalizovanih pogleda
- vrste s obzirom na vreme osvežavanja
 - **asinhrono (odloženo)**
 - u zadatom trenutku vremena
 - u regularnim vremenskim intervalima
 - na zahtev, prilikom izvođenja upita
 - na eksplicitni zahtev administratora
 - **sinhrono (trenutno)**
- vrste s obzirom na tehniku osvežavanja
 - **inkrementalno ("brzo") osvežavanje**
 - **kompletno osvežavanje**
- videti temu "Agregacija podataka u DW sistemima"

Osvežavanje podataka

Osvežavanje putem replikacije zasnovane na tokovima podataka

- **Data Streams tehnika**
- propagacija podataka, transakcija i događaja
 - unutar iste, ili između različitih baza podataka
- tri faze razmene tokova
 - **Capture**
 - detekcija izmena podataka i formiranje toka na izvoru
 - **Stage**
 - organizacija i memorisanje toka
 - **Consume (Apply)**
 - primena izmena – upotreba toka na cilju

Osvežavanje podataka

Osvežavanje putem replikacije zasnovane na tokovima podataka

- primer: Oracle tehnike Data Stream replikacije
 - **N-way replikacija**
 - multimaster replikacija tipa više-prema-više (n-to-n)
 - **Hub-and-Spoke replikacija**
 - jedan-prema-više (1-N) replikacija, od primarne (hub) lokacije prema udaljenim (spoke) lokacijama
 - **Table Replication with Synchronous Capture**
 - replikacija sa sinhronim preuzimanjem tokova
 - direktno preuzimanje izmena na osnovu sprovedenih DML naredbi
 - umesto preuzimanja podataka iz transakcijskog dnevnika
 - **Message Queuing with Streams Advanced Queuing (AQ)**
 - replikacija zasnovana na asinhronoj propagaciji poruka

Osvežavanje podataka

Osvežavanje zamenom kompletnih podataka

- celokupan stari sadržaj DW BP zamenjuje se novim
- jednostavna, ali generalno skupa tehnika
- ne razlikuje se postupak inicijalnog punjenja i osvežavanja DW BP
 - ekstreman slučaj inkrementalnog punjenja
 - koji se i ne može okarakterisati kao inkrementalno punjenje u pravom smislu značenja tog pojma
- primenljivo u slučaju malog obima DW BP
 - za pojedinačne DM sisteme
 - kada se ovakav postupak, zbog svoje jednostavnosti, više isplati od bilo koje druge tehnike

Osvežavanje podataka

Osvežavanje poređenjem instanci

- u odnosu na prethodno osvežavanje
- poređenjem instanci podataka, detektuju se razlike u sadržaju dve identično strukturirane BP
- razlike se evidentiraju u posebnoj tabeli ili fajlu
 - tzv. delta fajl
 - sadrži razlike podataka u odnosu na poslednje osvežavanje
 - delta fajl se obrađuje putem posebne procedure osvežavanja DW BP
- primenljivo u slučaju malog obima DW BP
 - kada je procedura poređenja dovoljno efikasna
 - tako da se ova tehnika više isplati od osvežavanja zamenom kompletnih podataka

Osvežavanje podataka

Osvežavanje sa vremenski obeleženim razlikama

- u tabelama izvora svaka izmena se vremenski označava i čuva se istorija ažuriranja
 - primenom neke od mogućih tehnika
 - uvođenjem posebnih "vremenskih" obeležja, verzija iste vrednosti ključa ili namenskih log (journal) tabela
- posebna procedura osvežavanja obrađuje vremenski označene podatke
 - započinjući od trenutka završetka prethodnog osvežavanja
- može se kombinovati sa tehnikom upotrebe trigera ili materijalizovanih pogleda

Osvežavanje podataka

Osvežavanje sa vremenski obeleženim razlikama

- Primer Oracle
 - Change Data Capture (CDC) Mechanism
 - paketi za publikovanje i preuzimanje izmena nad selektovanim izvornim tabelama
 - DBMS_CDC_PUBLISH
 - DBMS_CDC_SUBSCRIBE

Osvežavanje podataka

Osvežavanje upotrebom transakcionog dnevnika

- promene zabeležene u transakcionom dnevniku propagiraju se iz izvora u DW BP
 - postoji "before image" i "after image" svih ažuriranja
 - postoje kontrolni podaci o svakoj transakciji
- DW BP se osvežava u režimu oporavka
 - sprovođenja ažuriranja zasnovanog na sekvencijalnom čitanju transakcionog dnevnika
 - postoji vremensko označavanje podataka, oslonjenjo na operaciju pražnjenja svih bafera (checkpoint)
- često efikasna tehnika za osvežavanje

Osvežavanje podataka

Osvežavanje upotrebom transakcionog dnevnika

- naziva se i **transaction shiping**
- za razliku od prethodno pobrojanih tehnika koje podržavaju tzv. **data shiping**
- praktično, zahteva primenu identičnog SUBP na izvoru i u DW sistemu
 - takvog da podržava održavanje transakcionog dnevnika
 - teško je očekivati da je SUBP jednog proizvođača u stanju da interpretira interni format transakcionog dnevnika SUBP drugog proizvođača
- prethodne tehnike generalno ne uslovljavaju primenu identičnih SUBP na izvoru i u DW sistemu

Osvežavanje podataka

Politika osvežavanja

- definiše zahteve i opredeljenja u pogledu nabrojanih karakteristika osvežavanja, uvažavajući
 - zahteve i potrebe korisnika
 - kvalitet i karakteristike izvora i
 - tehnološke mogućnosti DW sistema
- zavisi od
 - definisanih zahteva i potreba korisnika, i prirode zadatka
 - intenziteta i obima ažuriranja podataka u izvorima
 - mogućnosti projektovane arhitekture DW sistema

Osvežavanje podataka

Politika osvežavanja

- zahtevi korisnika mogu rezultovati u potrebama
 - kratkog load window
 - npr. ako DW sistem obuhvata više vremenskih zona, te duži interval radnog vremena korisnika
 - više ili niže frekvencije osvežavanja
 - npr. satno / dnevno / nedeljno, ili čak
 - zahteva za trenutnim osvežavanjem DW BP
- visok intenzitet i obim ažuriranja podataka u izvorima \Rightarrow viša frekvencija osvežavanja
 - kako bi se smanjio obim procesiranih podataka i load window

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Arhiviranje i brisanje podataka

Arhiviranje

- neophodno zbog potreba
 - daljeg čuvanja starih podataka ili
 - obezbeđenja restauracije i eventualno oporavka DW sistema
- svodi se na raspoložive tehnike arhiviranja, podržane od strane SUBP, ili čak OS

Brisanje

- neophodno, bez obzira na dominantan istorijski karakter podataka u DW sistemima
- tehnike
 - TRUNCATE TABLE i DELETE FROM ...
 - ALTER TABLE ... [DROP | TRUNCATE | EXCHANGE] PARTITION ...

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Održavanje meta podataka

Kreiranje meta podataka

- u procesu razvoja DW sistema
- meta podaci predstavljaju projektnu dokumentaciju
- čuvaju se u repozitorijumu, u strukturiranom obliku

Održavanje meta podataka

- u ETL procesu
- prati održavanje podataka u DW BP

Administriranje meta podataka

- definisanje i uspostava odgovornosti nad meta podacima i procedurama, uključujući ETL procedure
 - definisanje uloga, prava pristupa i odgovornosti, korisnika
 - zaštita meta podataka i podataka – bezbednost i sigurnost

Održavanje meta podataka

Procesi nad meta podacima

- kreiranje, održavanje, administriranje
- u čvrstoj vezi sa grupom Configuration Management (CM) i Dependency Management procesa

Model strukture meta podataka

- model meta-meta nivoa
- OMG Common Warehouse Metamodel (CWM)
 - standardizovani model strukture meta podataka
 - cilj: obezbeđenje interoperabilnosti, tj. jednostavne razmene podataka između različitih alata koji su zasnovani na jedinstvenom CWM

Održavanje meta podataka

Globalni tipovi korisnika meta podataka

- **član razvojnog tima DW sistema**
 - projektanti, programeri, DBA
 - prava kreiranja i održavanja projektnih specifikacija
- **administrator DW sistema**
 - operativni administratori DW sistema
 - prava koja se odnose na
 - održavanje arhitekture DW sistema
 - sprovođenje ETL procesa
 - sprovođenje politike bezbednosti i sigurnosti sistema
- **krajnji korisnik DW sistema**
 - korisnici DW sistema, npr. menadžeri, u realizaciji svojih poslova
 - upotreba izveštajnih, OLAP ili Data Mining alata

Održavanje meta podataka

Meta podaci sadrže detaljne opise

- arhitekture DW sistema
 - raspoloživih računarskih i softverskih resursa DW sistema
- šeme DW BP
 - opisa činjeničnih i dimenzionih struktura
 - logičkih i fizičkih aspekata šeme BP
- izvora podataka
 - lokacija, načina nastanka i promena, opisa podataka
- DSA područja
 - opisa struktura podataka DSA područja i veza sa izvorima i DW sistemom

Održavanje meta podataka

Meta podaci sadrže detaljne opise

- ETL procesa
 - svih ETL procedura
 - pravila preslikavanja izvora u ciljne strukture
 - algoritama za transformaciju i agregaciju podataka
 - frekvencije, tehnika i procedura osvežavanja
- administrativnih procedura
 - koje se odnose na bezbednost, sigurnost i odgovornost nad podacima
 - koje omogućavaju praćenje realizacije svih zadataka u DW sistemu, tzv. auditing
- raspoloživih izveštaja
- raspoloživih metoda i alata za OLAP i Data Mining

Sadržaj

Procesi održavanja podataka u DW

Ekstrakcija podataka

Transformacija podataka

Punjenje podataka

Osvežavanje podataka

Arhiviranje i brisanje podataka

Održavanje meta podataka

Pitanja i komentari



Održavanje podataka u Data Warehouse sistemima

PREUZIMANJE, TRANSFORMACIJA I
PUNJENJE PODATAKA U DW BP

Kraj prezentacije