Upravljanje poslovnim podacima

Pročišćavanje podataka ETL

PROF. DRAŽENA GAŠPAR

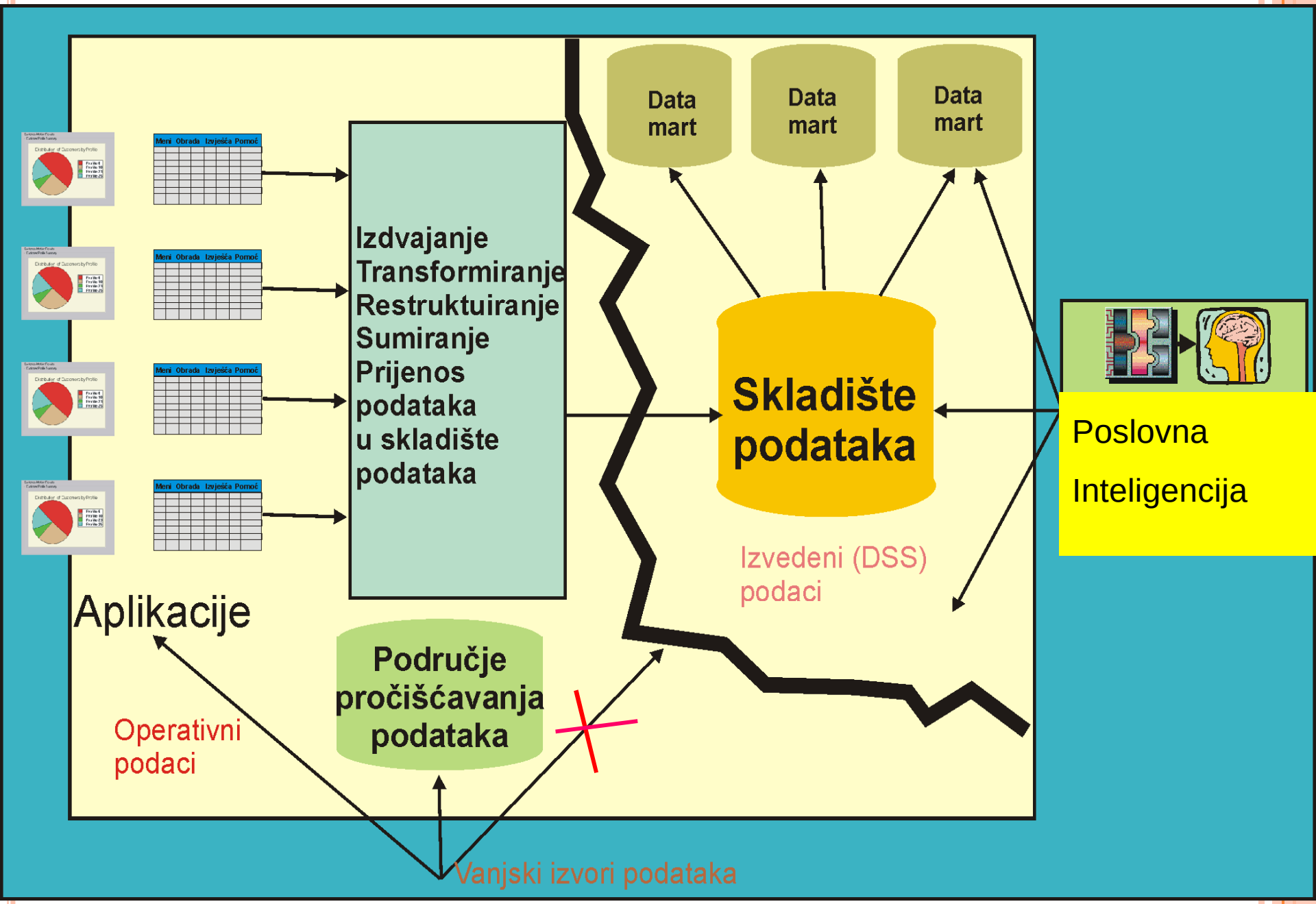
03.11.2015.

Prezentacije ..... Izvori podataka

Max. 5 minuta

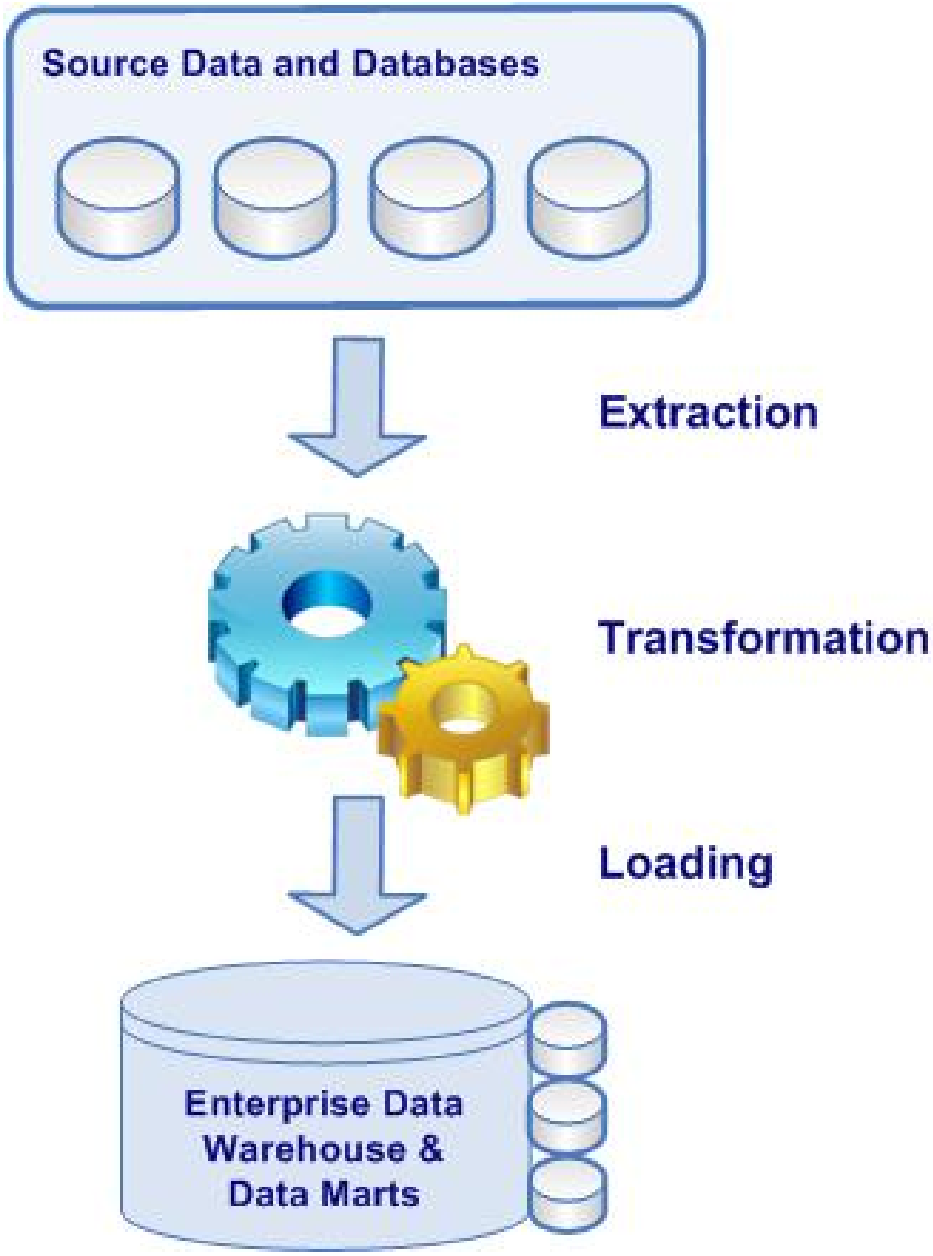
Pitanja za diskusiju

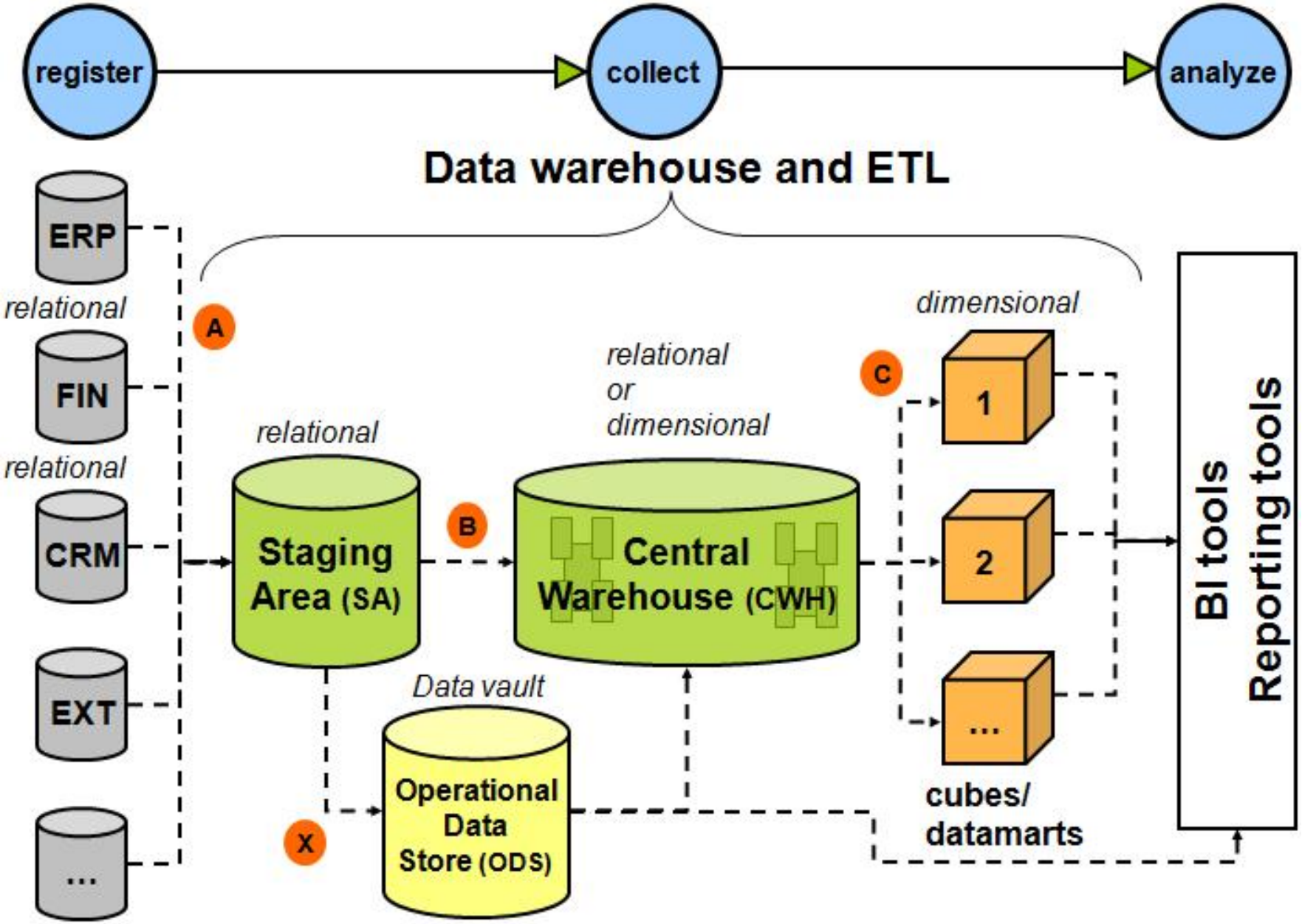
- Što se podrazumijeva pod pojmom ETL?
- Uz koju tehnologiju/je se veže pojam ETL?
- Koja je osnovna svrha ETL-a?
- Što su to “prljavi podaci”? Primjer!
- Koji su osnovni koraci ETL-a?
- Načini učitavanja/ažuriranja podataka?
- Koje su značajke kvalitetnog podatka?
- Gdje ste tražili informacije o ovome?


Data mart
Data mart
Data mart
Izdvajanje
Transformiranje
Restruktuiranje
Sumiranje
Prijenos
podataka
u skladište
podataka
Skladište
podataka
Poslovna
Inteligencija
Izvedeni (DSS)
podaci
Aplikacije
Područje
pročišćavanja
podataka
Operativni
podaci
Vanjski izvori podataka

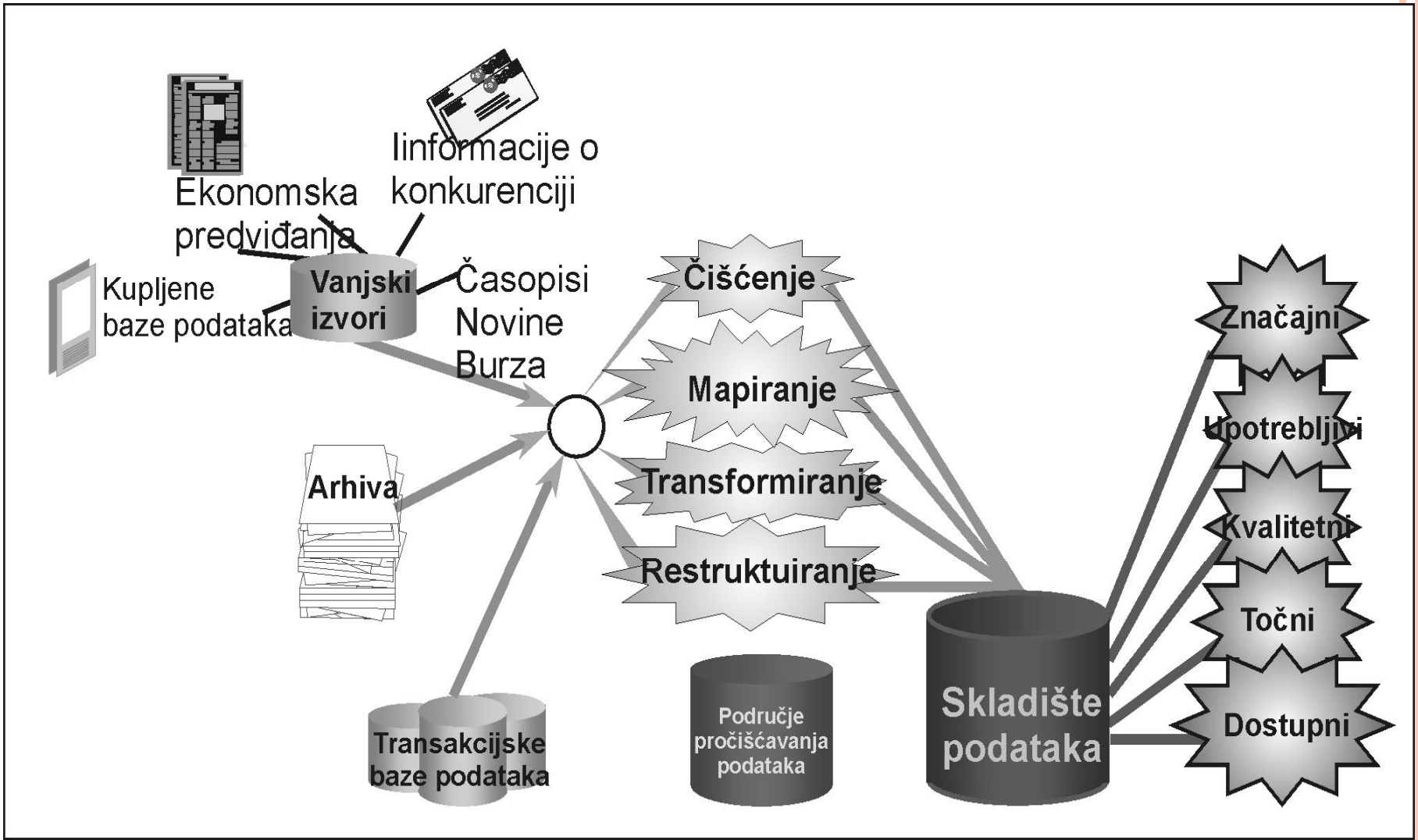
ETL

ETL


Source Data and Databases
Extraction
Transformation
Loading
Enterprise Data Warehouse & Data Marts


register
collect
analyze
Data warehouse and ETL
ERP
relational
FIN
relational
CRM
EXT
...
A
relational
Staging Area (SA)
B
relational or dimensional
Central Warehouse (CWH)
Data vault
X
Operational Data Store (ODS)
C
dimensional
1
2
...
cubes/ datamarts
BI tools
Reporting tools

Područje pročišćavanja podataka



Zašto je nužno pročišćavanje podataka?

- Podaci u realnosti su “prljavi”
  - nepotpuni: nedostaju vrijednosti atributa, neostaju određeni atributi ili sadrže samo agregirane podatke
    - npr. zanimanje=“ ”
  - sa šumovima: sadrže greške ili vrijednosti izvan granica
    - npr. Plaća=“-10”
  - nekonzistenti: sadrže neusklađenosti u kodovima ili nazivima/imenima
    - npr. Godine=“42” Datum rođenja=“03/07/1997”
    - npr. bilo je rangiranje “1,2,3”, sada je “A, B, C”
    - npr. neusklađenost duplih slogova

Problemi s integritetom podataka

- Ista osoba, različito napisano ime
  Dražena, Dražana, Draženka ...
- Više načina označavanja naziva kompanije
  - Hera, Hera d.o.o., Hera – SW kompanija
- Uporaba različitih naziva
  Mumbai, Bombay
- Različite šifre generirane od strane različitih aplikacija za istog kupca
- U obvezna polja unešen znak blank, . i sl.
- Pogrešna šifra proizvoda unešena na POS-u
  - Ručni unosi dovode do grešaka
  - “u slučaju problema koristiti 9999999”

Zašto su podaci “prljavi”?

- Nepotpuno podaci mogu proizići iz
  - “Nije primjenjivo” kao podatkovna vrijednost kod prikupljanja
  - Razlika u razmišljanju u vrijeme kada su podaci prikupljani i kada se analiziraju.
  - Ljudski/hardware/software problemi
- Šumovi (netočne vrijednosti) podataka mogu proizići iz
  - Neispravni uređaji prikupljanja podataka
  - Ljudske ili računalne greške na unosu podataka
  - Greške u prijenosu podataka
- Nekonzistentnost podataka može proizići iz
  - Različiti izvori podataka
  - Narušavanje funkcijskih ovisnosti (npr. Izmjene povezanih podataka)
- Dupli slogovi također traže čišćenje

Zašto je nužno pročišćavanje podataka?

- Bez kvalitetnih podataka nema ni kvalitetnih rezultata analize!
  - Kvalitetne odluke morju biti temeljene na kvalitetnim podacima
    - Npr. Dupliciranі ili nedostajući podaci mogu uzrokovati pogrešnu ili varljivu statistiku.
  - Skladište podataka treba konzistentnu integraciju kvalitetnih podataka
  - Proces izdvajanja, čišćenja i transformiranja podataka čini najveći dio posla izgradnje skladišta podataka

Višedimenzijska mjera kvalitete podataka

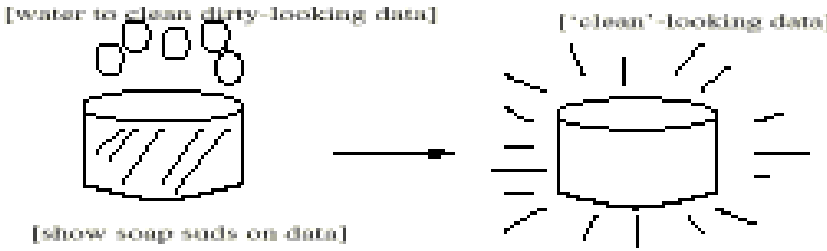
- Opće prihvaćeni višedimenzijski pogled:
  - Točnost
  - Potpunost
  - Dosljednost (konzistentnost)
  - Pravovremenost
  - Vjerodostojnost
  - Dodatna vrijednost
  - Interpretativnost
  - Pristup
- Opće kategorija:
  - Značajnost, odgovara kontekstu, reprezentativan, pristupačan

Osnovni zadaci

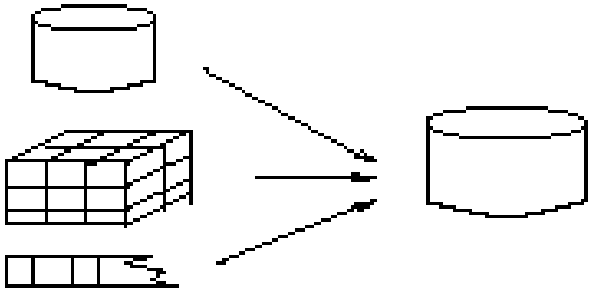
- Čišćenje podataka
  - Popunjavanje vrijednosti koje nedostaju, rješavanje šumova, identificiranje ili uklanjanje “nepodobnih” i razrješavanje nekonzistentnosti
- Integriranje podataka
  - Integriranje višestrukih baza podataka, podatkovnih kocki ili datoteka
- Transformiranje podataka
  - Normalizacija i agregacija
- Smanjivanje (redukcija) podataka
  - Smanjivanje obima podatka uz zadržavanje istih ili sličnih analitičkih rezultata
- Diskretizacija podataka
  - Dio smanjivanja (redukcije) podataka, ali posebice bitno za numeričke podatke

PRIMJERI OSNOVNIH ZADATAKA

Data Cleaning



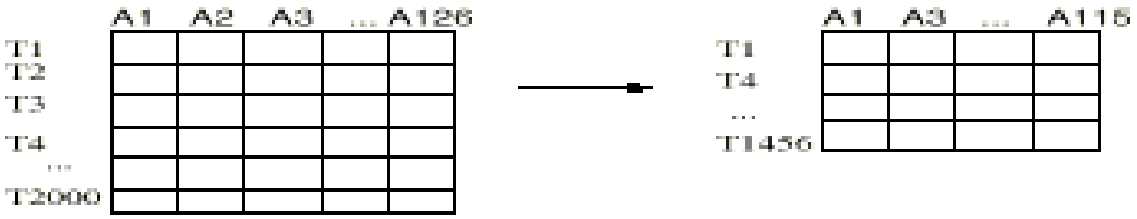
Data Integration



Data Transformation

-2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Čišćenje podataka

- Rješavanje
  - Nepotpunih podataka
  - Šumova u podacima
  - Nekonzistentnosti

Nepotpuni (nedostajući) podaci

- Podaci nisu raspoloživi
  - Npr. Mnoge n-torke nemaju zabilježene vrijednosti za neke od atributa (null polja)
- Podaci mogu nedostajati zbog:
  - Grešaka na opremi
  - Nekonzistentnosti s drugim podacima, pa su zbog toga izbrisani
  - Podaci nisu unešeni zbog nerazumijevanja
  - Određeni podaci nisu smatrani bitnim u vrijeme unosa
  - Ne prati se povijest promjena nad podacima

ŠTO RADITI S NEPOTPUNIM PODACIMA?

- Ignorirati n-torku: obično se radi kada nedostaje oznaka klase (kada se radi klasifikacija)—nije efektivno kada postotak nedostajućih vrijednosti značajno varira po atributima
- Ručno popunjavanje nedostajućih vrijednosti: zamorno + neizvodljivo?
- Automatsko popunjavanje s:
  - Općom konstantom: npr. “nepoznato”, nova klasa?!
  - Srednja vrijednost atributa
  - Srednja vrijednost atributa za sve koji pripadaju istoj klasi: mudriji pristup
  - Najvjerojatnija vrijednost: zaključak temeljen na Bayesian formuli ili stablu odlučivanja

Šumovi u podacima (Noisy Data)

- Šum: random greška ili promjena (odstupanje) u mjerenju
- Netočne vrijednosti atributa zbog:
  - Manjkavih instrumenata prikupljanja podataka
  - Problema na unosu podataka
  - Problema pri slanju podataka
  - Tehnoloških ograničenja
  - Nekonzistentnosti u imenovanju
- Drugi problemi koji traže čišćenje podataka su:
  - Dupli slogovi
  - Nepotpuni podaci
  - Nekonzistentni podaci

ŠTO RADITI SA ŠUMOVIMA U PODACIMA?

- Binning (grupiranje)
  - Prvo se podaci sortiraju i podijele u (jednake frekfencije) grupe
  - Onda se radi “gla đenje” (izravnavanje) grupa pomoću srednje vrijednosti, medijane, graničnih vrijednosti i sl.
- Regresija
  - Izravnavanje pomoću regresijskih funkcija
- Klasteriranje
  - Otkrivanje i otklanjanje nepodobnih članova grupe
- Kombiniranje računalne i ljudske kontrole
  - Otkrivanje sumnjivih vrijednosti i provjera od strane ljudi (npr. Posebno kod nepodobnih članova grupe)

Čišćenje podataka kao proces

- Otkrivanje proturječnosti u podacima
  - Uporaba metapodataka (npr. domene, rang, ovisnost, distribucija)
  - Provjera prekoračenja polja
  - Provjera pravila jedinstvenosti, dosljednosti i nul vrijednosti
  - Uporaba komercijalnih alata
    - Čišćenje podataka (data scrubbing): uporaba jednostavnog znanja o domenama (npr. Poštanski broj, spell-check) za otkrivanje grešaka i korekcije
    - Revizija podataka (data auditing): kroz analizu podatakak kako bi se otkrila pravila i veze za za pronalaženje onih koji krše pravila (npr. Korelacija i klasteriranje za pronalaženje nepodobnih)
- Migriranje i integriranje podataka
  - Alati za migriranje podataka: omogućavaju specificiranje transformacija
  - ETL alati: omogućavaju korisnicima specificiranje transformacija kroz GUI
- Integriranje dva procesa
  - Iterativnog i interaktivnog

Integriranje podataka

- Integriranje podataka:
  - Kombiniranje podataka iz višestrukih izvora u koherentnu bazu
- Shema integracija: npr. A.cust-id ≡ B.cust-#
  - Integriranje metapodataka iz različitih izvora
- Problem identificiranja entiteta:
  - Identificiranje stvarnih entiteta iz višestrukih izvora podataka, npr. Bill Clinton = William Clinton
- Otkrivanje i rješavanje konfliktnih vrijednosti podataka
  - Za isti stvarni entitet, vrijednosti atributa su različite za različite izvore
  - Mogući razlozi: različito predstavljanje, različite skale, npr. Metrička vs Britanska

Redundancija i integriranje podataka

- Redundantni podaci se javljaju najčešće kada se radi integriranje iz više baza podataka
  - *Identifikacija objekta*: Isti atribut ili objekt mogu imati različita imena u različitim bazama
  - *Izvedeni podaci:* Jedan atribut može biti "izvedeni" atribut u drugoj tablici, npr. Godišnji prihod
- Redundantni atributi se mogu otkriti korelacijskom i analizom kovarijanse
- Pažljivo integriranje podataka iz više izvora može pomoći u smanjenju/izbjegavanju redundanse i nekonzistentnosti, te poboljšati brzinu rudarenja i kvalitetu podataka

23

Transformiranje podataka

- Izravnjavanje (smoothing): uklanjanje šuma u podacima
- Agregiranje: zbrajanje, izrada podatkovne kocke
- Generalizacija: koncept “penjanja” po hijerarhiji
- Normalizacija: promjena veličine kako bi se uklopilo unutar manjeg, specifičnog ranga
  - min-max normalizacija
  - z-score normalizacija
  - normalization pomoću decimalnog skaliranja
- Konstruiranje atributa/osobina
  - Novi atributi nastaju iz postojećih

cept
s

Transformiranje podataka: Normalizacija

- Min-max normalizacija: za[new_min$_A$, new_max$_A$]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Npr. Neka se prihod s ranga $12,000 do $98,000 normalizira na [0.0, 1.0]. Onda se $73,000 mapira na $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score normalizacija (μ: srednja vrijednost, σ: standardna devijacija):

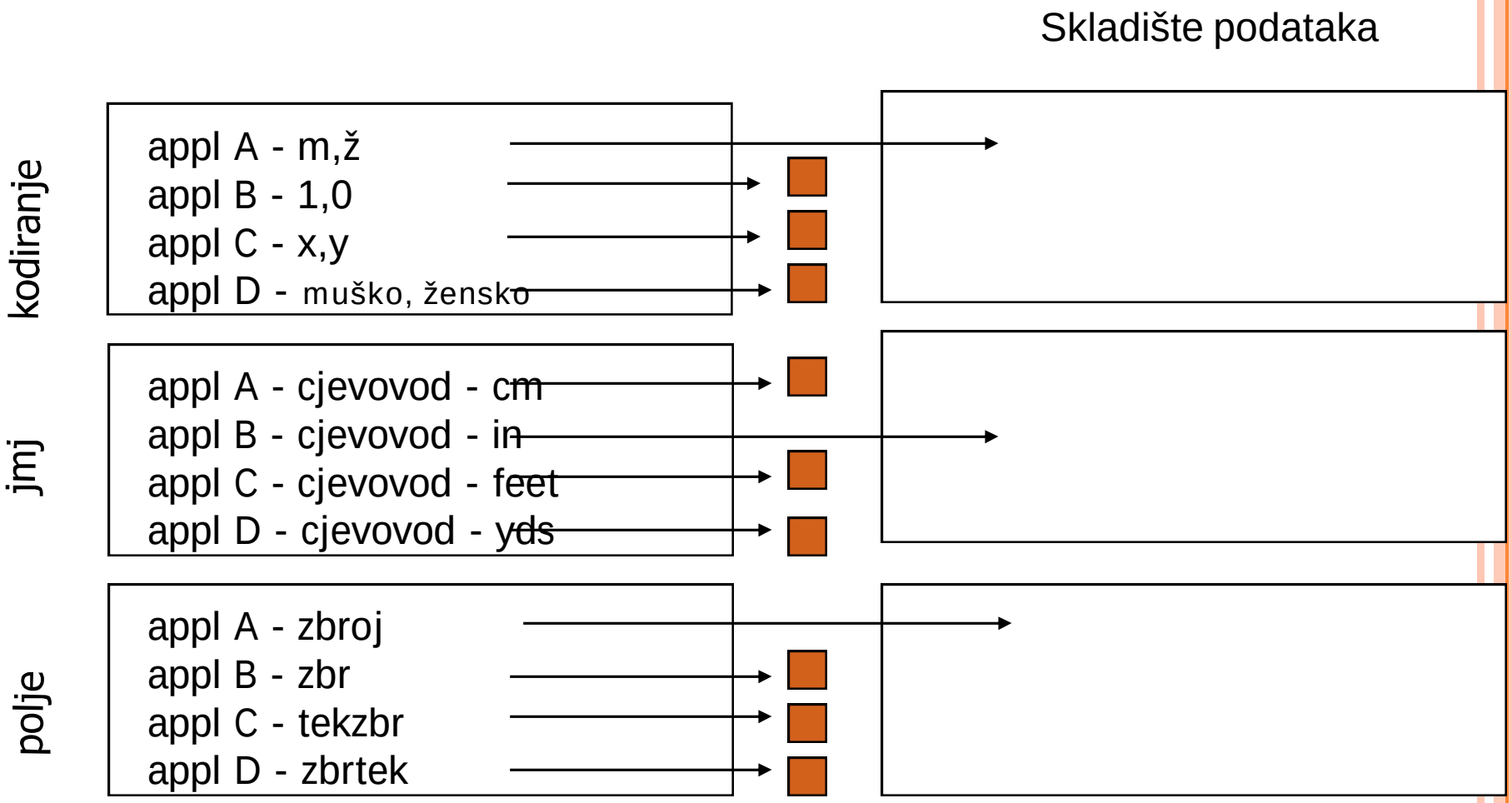
$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Npr. Neka je μ = 54,000, σ = 16,000. Slijedi $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- Normalizacija pomoću decimalnog skalirnaja

$$v' = \frac{v}{10^j}$$ Gdje je j najmanji cio broj takav da je Max(|v'|) < 1

Transformiranje podataka - Primjer



Strategije redukcije podataka

- Redukcija podataka: dobiti reducirano predstavljanje skupa podataka koji je znatno manji po veličini ali daje iste (ili skoro iste) analitičke rezultate
- Zašto redukcija podataka? — Baze/skladišta podataka mogu pohranjivari terabyte podataka. Složena analiza podataka može trajati vrlo dugo ako se izvršava na potpunom skupu podataka.
- Strategije redukcije podataka
  - Dimenzijska redukcija, npr. Uklanjanje nevažnih atributa
    - Elementarne transformacije
    - Principal Components Analysis (PCA)
    - Odabir podskupa mogućnosti, kreiranje mogućnosti
  - Smanjenje veličine (neki to zovu i redukcija podataka)
    - Regresija i Log-Linear Models
    - Histogrami, klasteriranje, uzorci
    - Gregiranje podatkovnih kocki
  - Kompresija podataka

Područje pročišćavanja podataka

ETL (Extraction Transformation Loading)

- Otkrivanje promjena u izvornim podacima potrebnim za skladište podataka;
- Izdvajanje podataka iz izvornih sustava;
- Čišćenje i transformiranje podataka;
- Restrukturiranje ključeva podataka;
- Indeksiranje podataka;
- Sumiranje podataka;
- Održavanje metapodataka;
- Učitavanje podataka u skladište podataka.

Pojmovi transformiranja podataka

- Izdvajanje
- Prilagodba
- Ribanje
- Miješanje
- Kućanstva
- Obogaćivanje
- Procjenjivanje
- Učitavanje
- Validacija
- Ažuriranje

Pojmovi transformiranja podataka

- Izdvajanje
  - Izdvaja podatke iz operativnih izvoru u “as is” statusu
- Prilagodba
  - Konverzija tipova podataka iz izvornih u krajnje baze (skladište podataka)

Data Transformation Terms

- Kućanstva
  - Identificiranje svih članova kućanstva (koji žive na istoj adresi)
  - Osigurava da se samo jedna poštanjska pošiljka šalje kućanstvu
  - Može rezultirati značajnim uštedama papira, poštarine …

Data Transformation Terms

- Obogaćivanje
  - Korištenje podataka iz vanjskih izvora kako bi se obogatili operativni podaci.

- Procjenjivanje
  - Izračun vjerojatnosti događaja

    npr. Vjerojatnost da će kupac kupiti novi proizvod, promijeniti marku proizvoda

Učitavanje (Load)

- Nakon izdvajanja, ribanja, čišćenja, validiranja itd. potrebno je učitati podatke u skladište podataka

- Otvorena pitanja
  - Ogromne količine podataka koje treba učitati
  - Kratko vrijeme kada skladište podataka može biti off line (često ne ni noću - web)
  - Kada praviti indekse i zbrojne tablice
  - Dozvoliti administratoru sustava nadzor, prekid, nastavak, promjenu stope učitavnja
  - Skladan oporavak – nastavak nakon ispada sustava tamo gdje se stalo bez gubitka integriteta podataka

Tehnike učitavanja

- Korištenje SQL-a za dodavanje ili unos novih podataka
  - Slog u određenom vremenu
  - Dovodi do random disk I/O
- Korištenje batch učitavanja

Taksonomija učitavanja

- Inkrementalni naspram potpunog učitavanja
- Online naspram Offline učitavanja

Ažuriranje (osvježavanje)

- Propagira ažuriranja nad izvornim podacima na skladište podataka

- Otvorena pitanja:
  - Kada osvježiti (refresh)
  - Kako osvježiti – tehnike ažuriranja

Kada osvježiti?

- periodično (npr. svaku večer, svaki tjedan) ili nakon značajnih događaja
- Za svako ažuriranje: nije zajamčeno sve dok DW ne zatraži ažuran podatak
- Politika osvježavanja postavljena od strane administratora a bazirana na korisničkim potrebama i prometu
- Moguće različite politike za različite izvore podataka

Tehnike osvježavanja

- Potpuno izdvajanje iz osnovnih tablica
  - Čita čitavu izvornu tablicu: preskupo
  - Možda jedini izbor za nasljeđene sustave

Kako otkriti promjene

- Kreirati snapshot log tablicu za bilježenje id-ijeva ažuriranih redaka izvornih podataka i timestamp-ova

- Otkrivanje promjena pomoću:
  - Definiranja “after row” okidača (triggers) za ažuriranje snapshot loga kada se promijeni izvorna tablica
  - Korištenje regularnih transakcijskih logova za otkrivanje promjena u izvornim podacima

Izdvajanje podataka i čišćenje

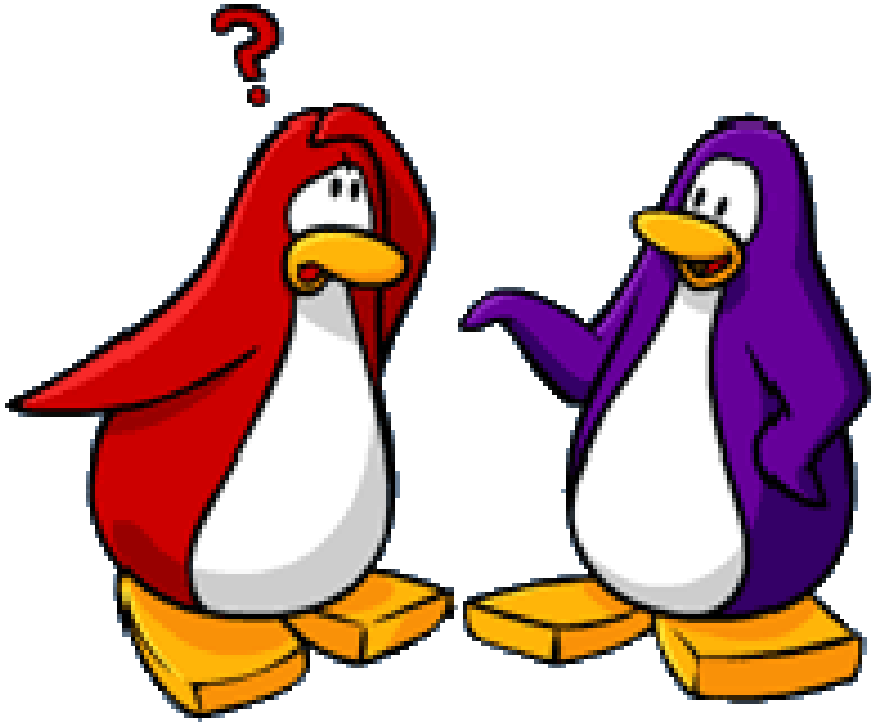
- Izdvajanje podataka iz postojećih operativnih i nasljeđenih podataka
- Otvorena pitanja:
  - Izvori podataka za DW
  - Kvaliteta izvornih podataka
  - Miješanje različitih izvora podataka
  - Transformiranje podataka
  - Kako propagirati ažuriranja (na izvornim podacima) u skladište podataka
  - Terabytes podataka za učitavanje

Za sljedeće predavanje

- Datum: 10.11.2015.

1. Prezentirati:
   - Opis problema
   - ER model podataka
   - Use case dijagram
   - Dijagram aktivnosti

Prezentaciju i MS Word dokument s detaljnim opisom i svim dijagramima poslati prof. Gašpar na mail NAJDALJE DO 09.11.2015., 23:59



Questions..