Skladištenje podataka

Kimballova tehnička arhitektura i ETL proces

Prof.dr.sc. Dražena Gašpar

14.11.2016.

Prezentacije .....

Opis problema +

ER model i izvori podatka
- minimalno 1 vanjski izvor

Max. 5 minuta

Arhitektura

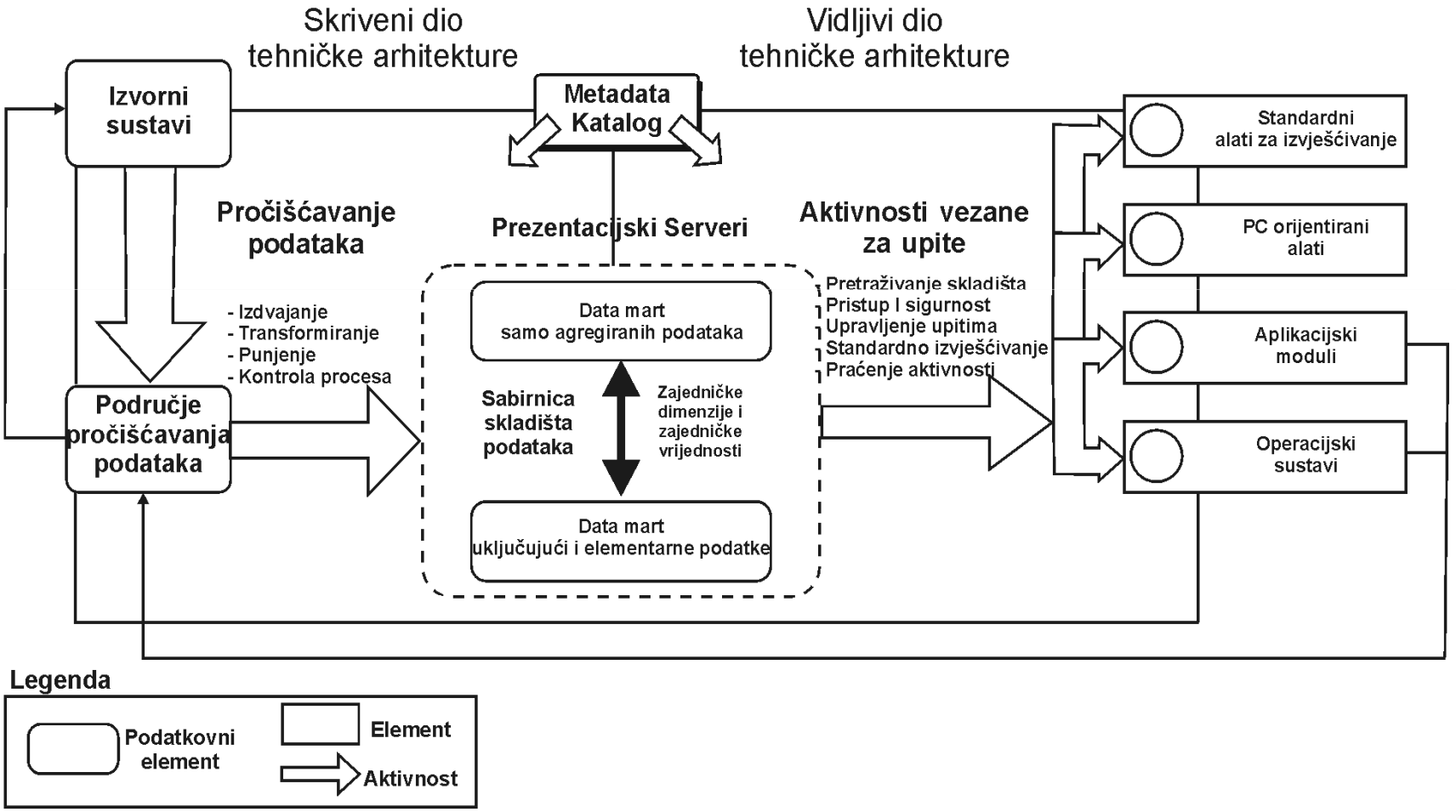
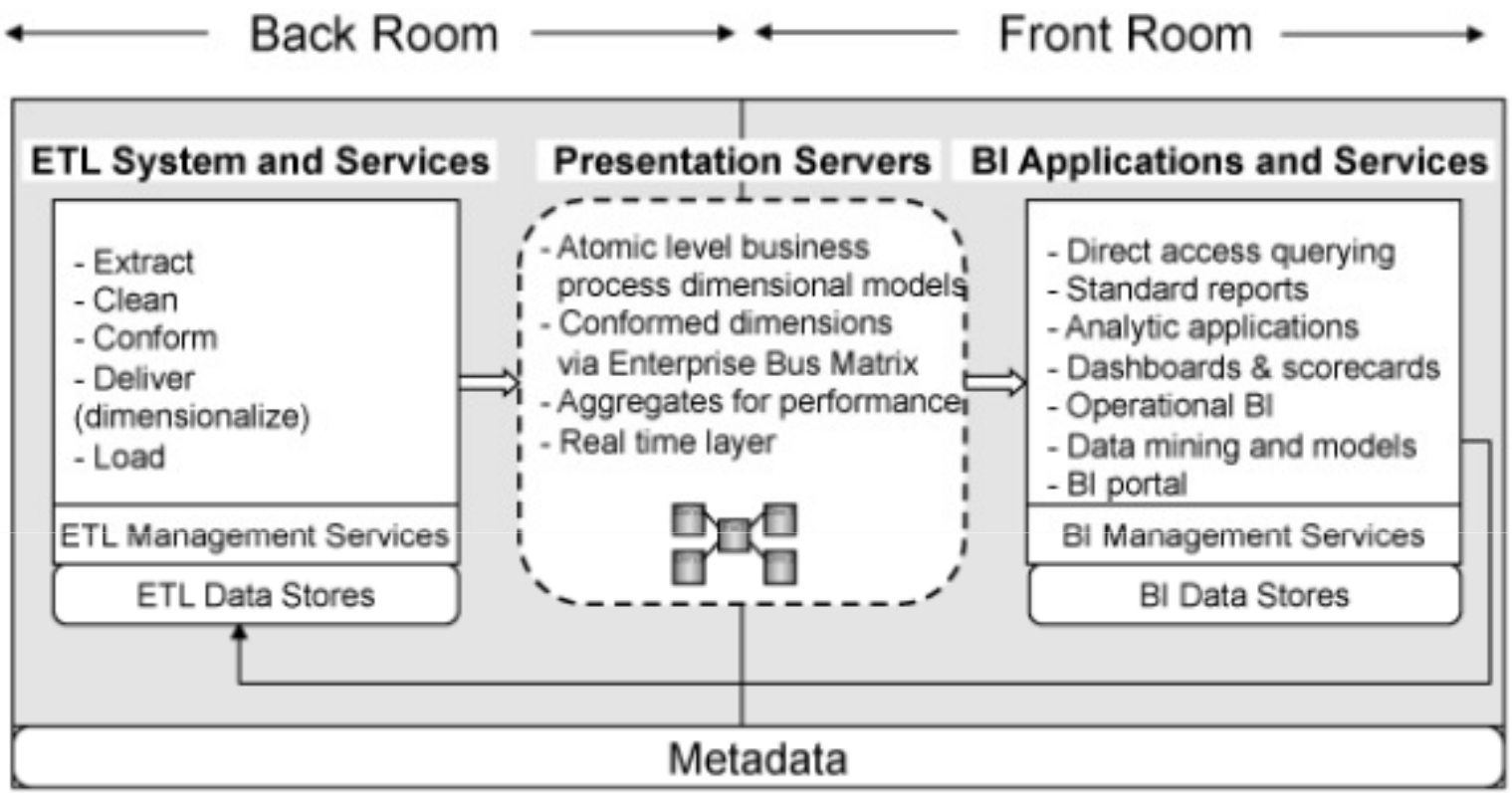
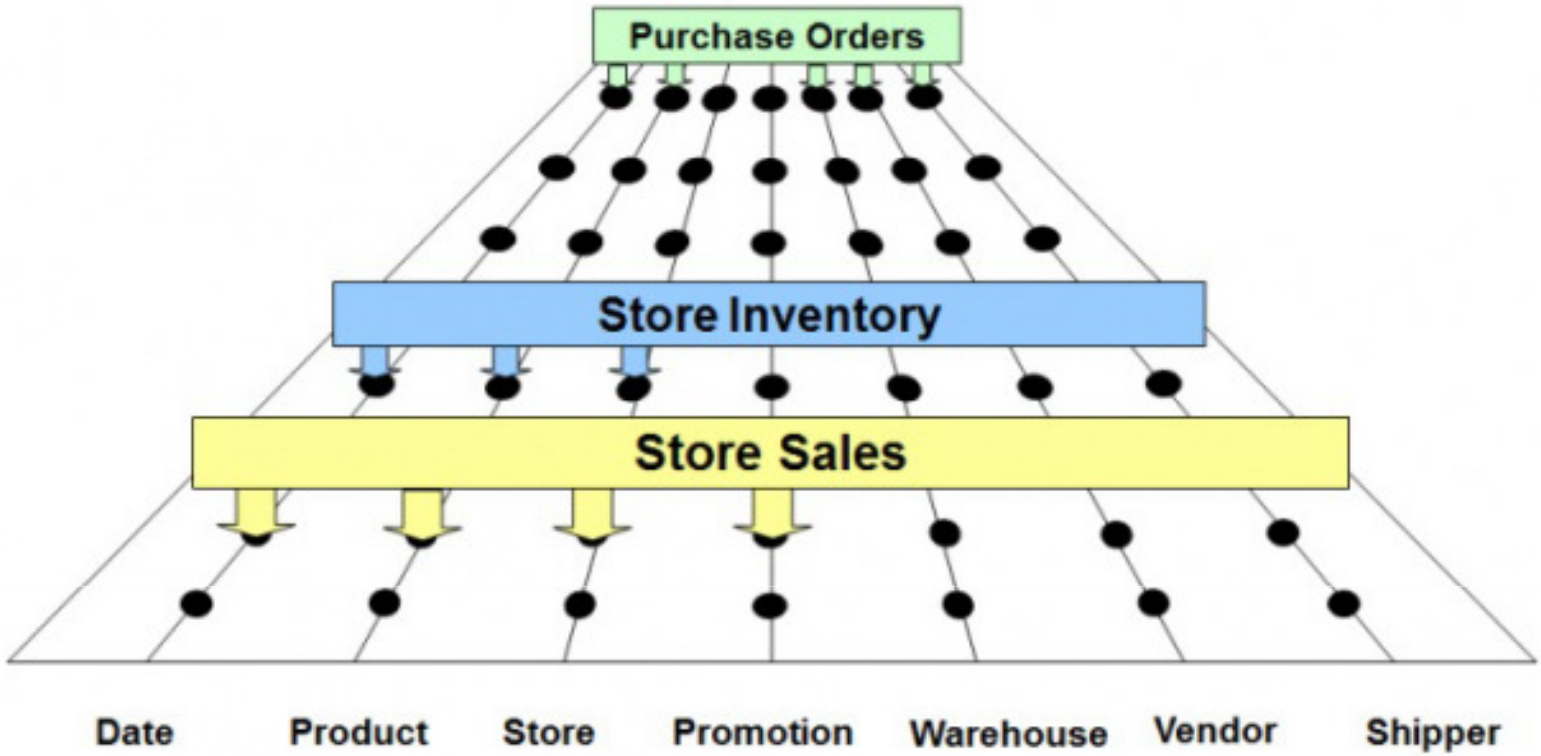


Figure 1: Kimball technical system architecture diagram.

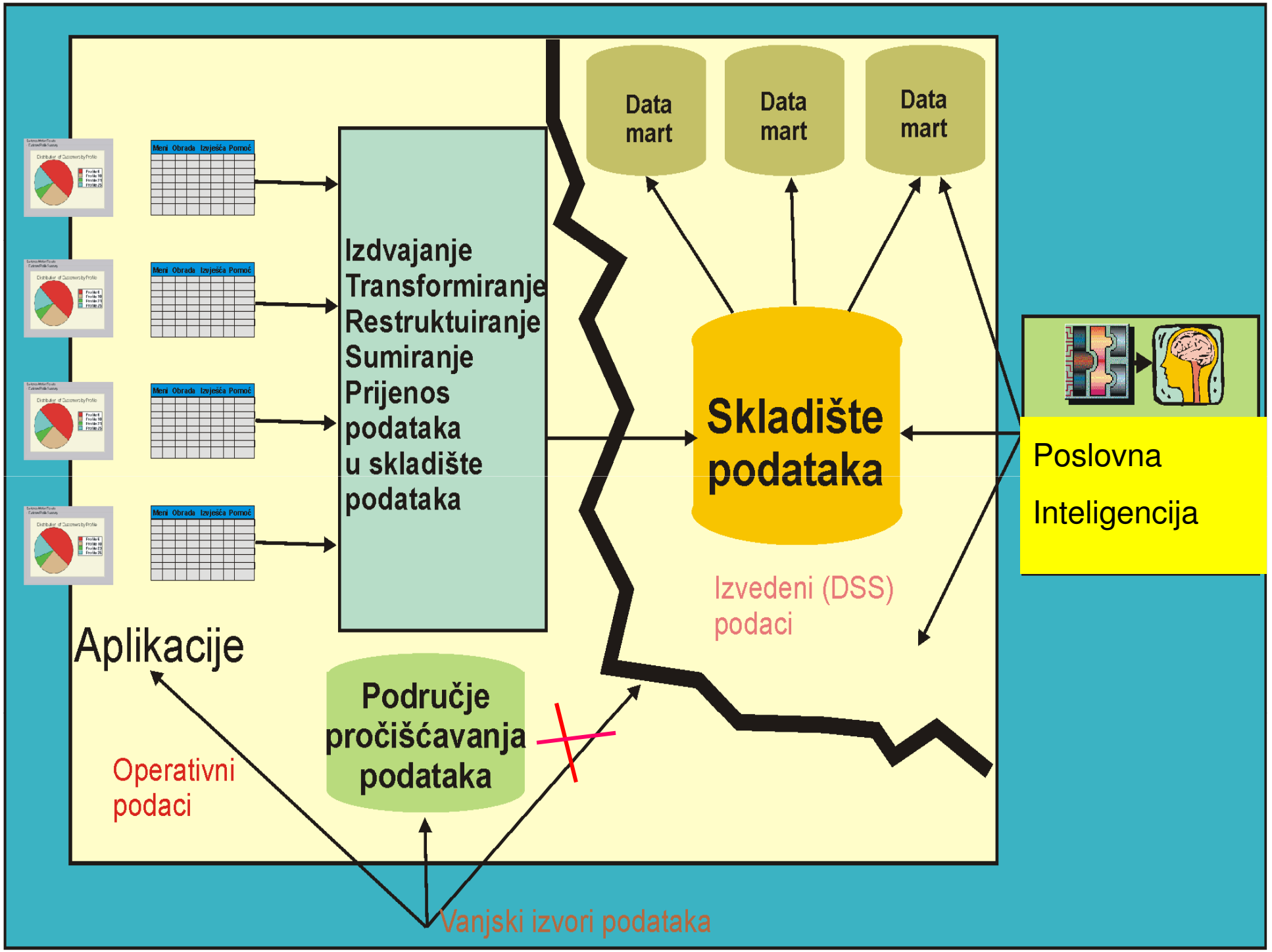


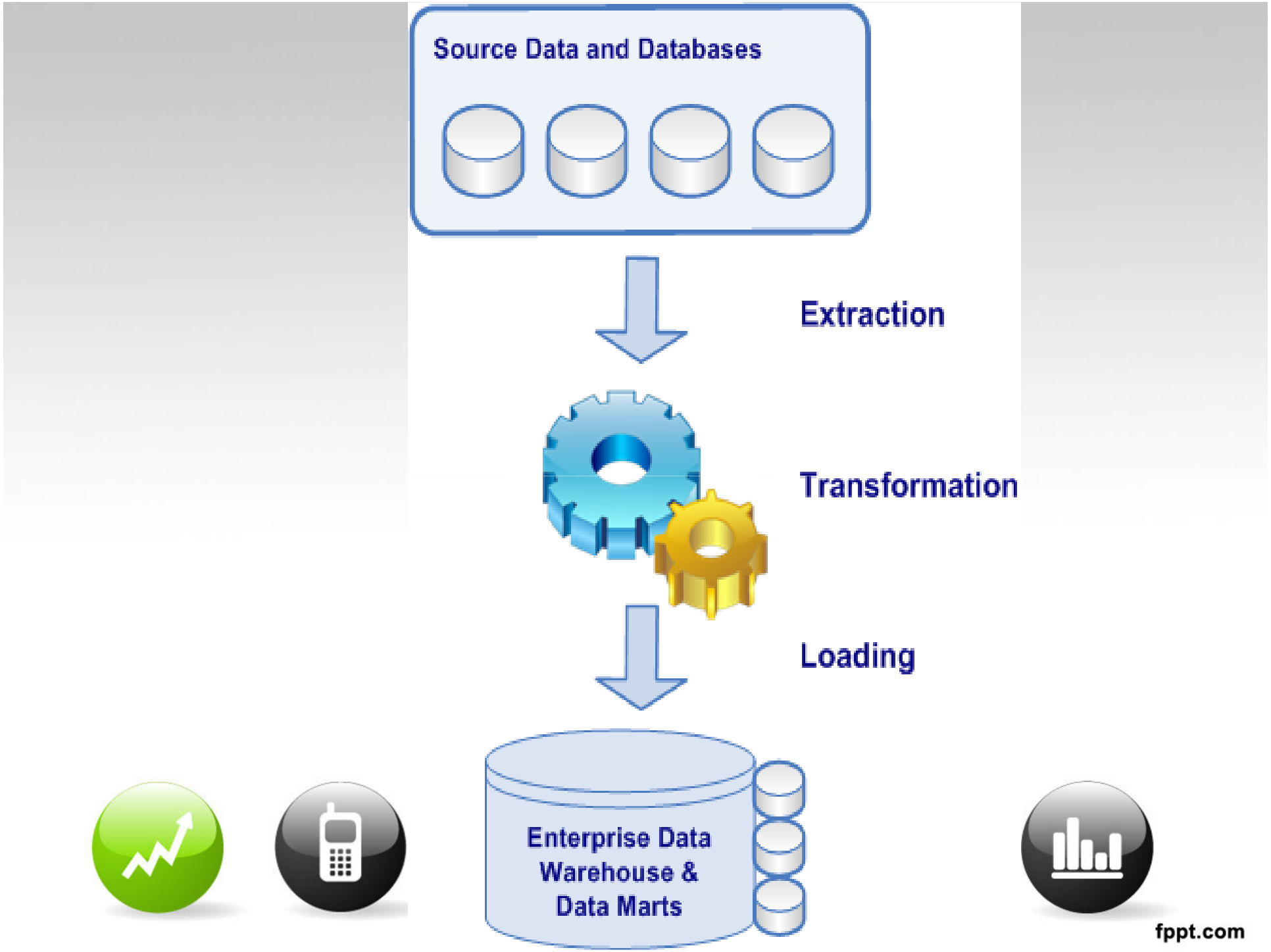
Source: file:///C:/Users/user/Downloads/kimballgroup.com-Kimball%20Technical%20DWBI%20System%20Architecture.pdf

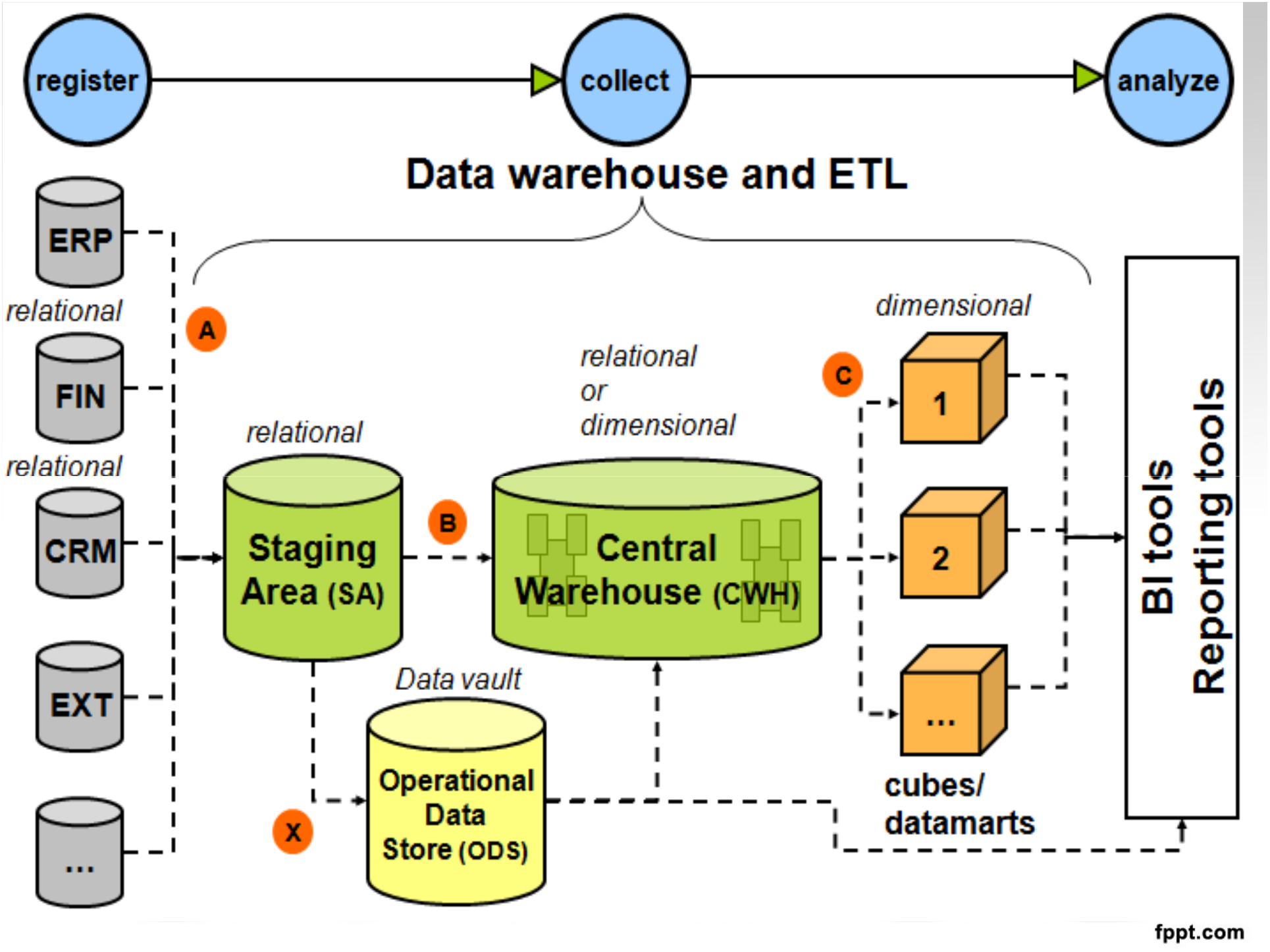


Source: file:///C:/Users/user/Downloads/kimballgroup.com-Enterprise%20Data%20Warehouse%20Bus%20Architecture.pdf

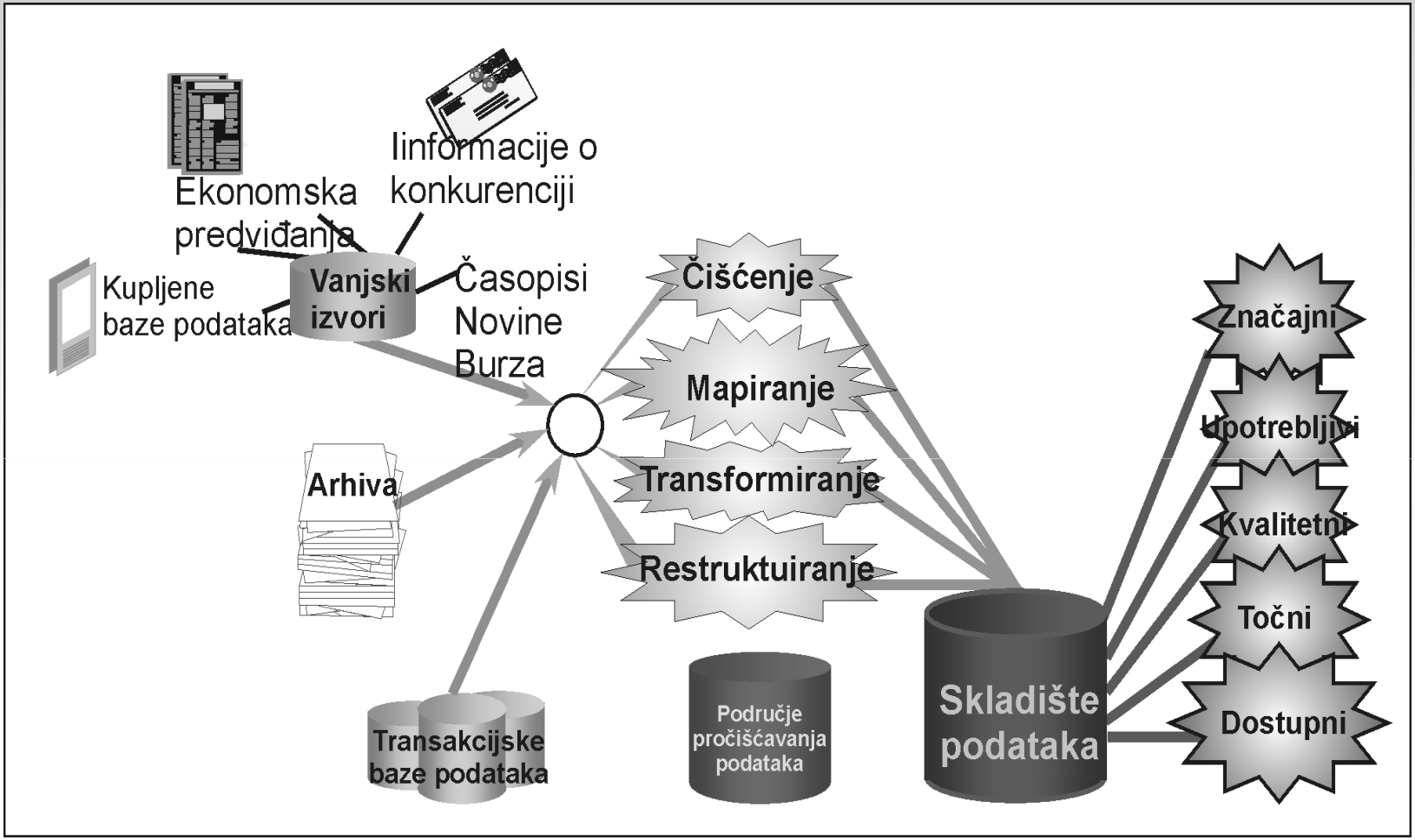
ETL


Data mart
Data mart
Data mart
Meni Obrada Izvješća Pomoć
Meni Obrada Izvješća Pomoć
Meni Obrada Izvješća Pomoć
Meni Obrada Izvješća Pomoć
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


Source Data and Databases
Extraction
Transformation
Loading
Enterprise Data Warehouse & Data Marts
fppt.com


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
fppt.com

ETL - 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 razičitih 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 use 9999999”

Zašto su podaci “prljavi”?

- Nastajanje nepotpunih podataka:
  - “Not applicable” vrijednost podatka u trenutku prikupljanja
  - Različito poimanje u vrijeme prikupljanja i analize podataka.
  - Ljudski/hardware/software problemi
- Nastajanje šumova u podacima (pogrešne vrijednosti):
  - Pogreške na istrumentima za prikupljanje podataka
  - Ljudske ili računalne greške na unosu podataka
  - Greške pri prijenosu podataka
- Nastajanje nekonzistentnih podataka:
  - Različiti izvori podataka
  - Narušavanje funkcionalne ovisnosti (izmjena povezanih podataka)
- Duple slogove također treba “očistiti”

Zašto je nužno pročišćavanje podataka?

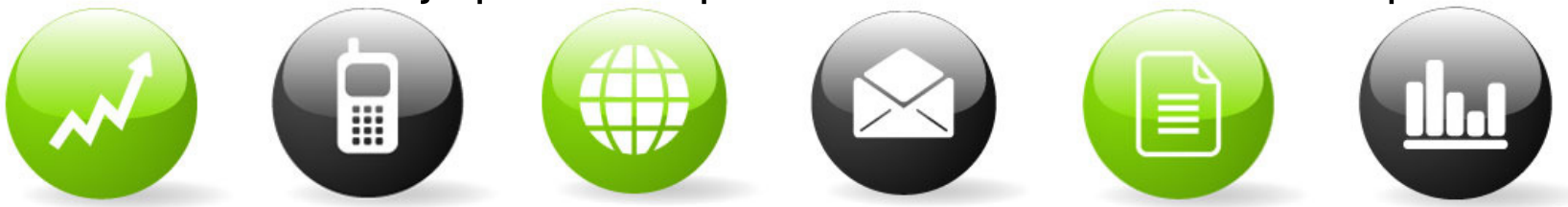
- Bez kvalitetnih podataka, nema kvalitetnih rezultata data mining-a!
  - Kvalitetno odlučivanje se mora temeljiti na kvalitetnim podacima
    - npr., dupli ili nedostajući podaci mogu dovesti do pogrešne ili čak zbunjujuće statistike.
  - Skladište podataka treba konzistentnu integraciju kvalitetnih podataka
- ETL proces čini većinu posla pri razvoju skladišta podataka

Višedimenzijsko mjerenje kvalitete podataka

- Općepihvaćeni višedimenzijski pristup:
  - Točnost (Accuracy)
  - Potpunost (Completeness)
  - Konzistentnost (Consistency)
  - Pravovremenost (Timeliness)
  - Vjerodostojnost (Believability)
  - Dodatna vrijednost (Value added)
  - Interpretativnost (Interpretability)
  - Dostupnost (Accessibility)
- Šire kategorije:
  - Suštinski, odgovara kontekstu, reprezentativan i dostupan

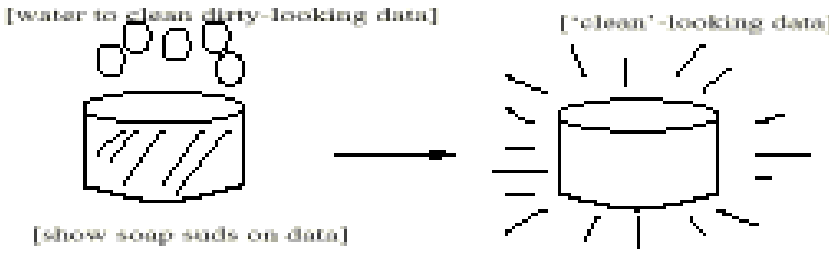
Osnovni ETL zadaci

- Čišćenje podataka (Data cleaning)
  - Popunjavanje nedostajućih vrijednosti, rješavanje šumova u podacima, pronalaženje i uklanjanje nepodobnih članova grupe, rješavanje nekonzistentnosti
- Integriranje podataka (Data integration)
  - Integriranje više baza podataka, podatkovnih kocki ili datoteka
- Transformiranje podataka (Data transformation)
  - Normalizacija i agregacija
- Reduciranje podataka (Data reduction)
  - Formiranje reduciranih setova podataka koji daju iste ili slične analitičke rezultate kao da se analiziraju svi podaci
- Diskretizacija podataka (Data discretization)
  - Dio reduciranja podataka posebice važan za numeričke podatke

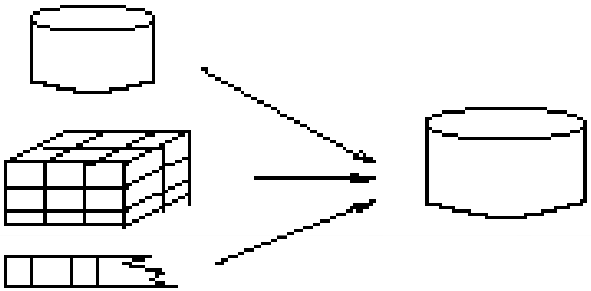


Oblici ETL-a

Data Cleaning



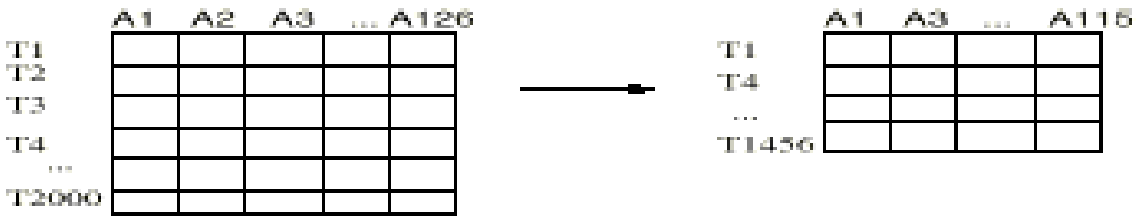
Data Integration



Data Transformation

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

Data Reduction



Deskriptivno sumiranje podataka

- Motivacija
  - Bolje razumijevanje podataka: centralna tendencija, varijacija i disperzija
- Značajke disperzije podataka
  - median, max, min, izvan granica, varijansa, itd.
- Numeričke dimenzije oodgovaraju sortiranim intervalima
  - Disperzija podataka: analiziranje različite usitnjenosti i preciznosti
- Analiza disperzije na izračunatim mjerama
  - Promjena od početne do krajnje vrijednosti kod numeričkih podataka

Mjere centralne tendencije

- Aritmetička sredina (uzorak vs. populacija): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$
  - Ponderirana aritmetička sredina:
  - Skraćena srednja vrijednost: izbacivanje ekstrema $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Medijan: Potpuna mjera
  - Srednja vrijednost za neparan broj vrijednosti ili prosjek dvije srednje vrijednosti

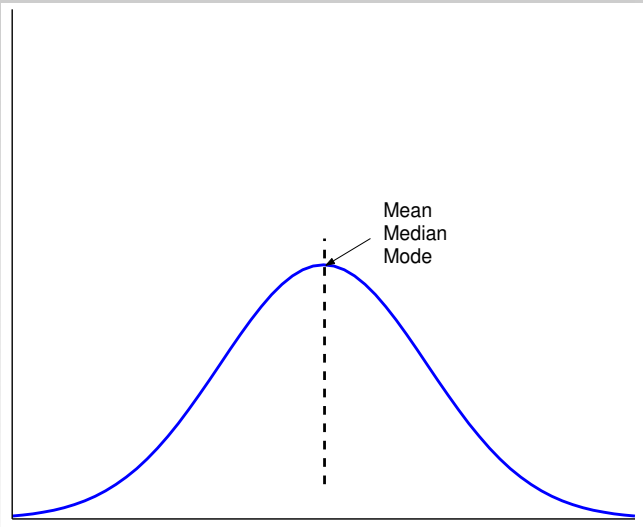
$$median = L_1 + (\frac{n/2 - (\sum f)l}{f_{median}})c$$

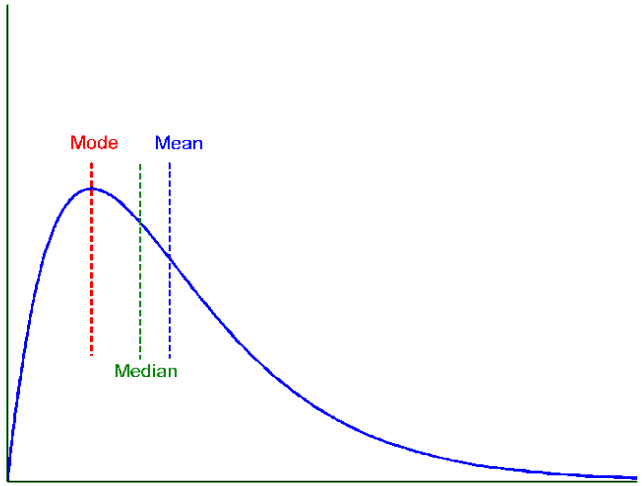
- Mod
  - Vrijednost koja se najčešće pojavljuje u podacima
  - Unimodal, bimodal, trimodal
  - Empirijska formula:

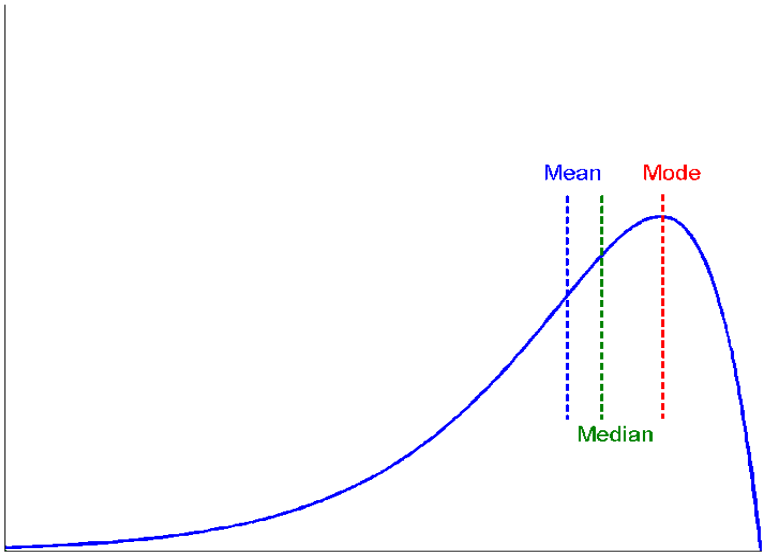
$$mean - mode = 3 \times (mean - median)$$

Simetrični vs. Asimetrični podaci

- Medijan, aritmetička sredina i mod za simetrične, pozitivno i negativno asimetrične podatke







Čišćenja podataka

- Značaj
  - “Čišćenje podatka je jedan od tri najveća problema u skladištenju podataka”—Ralph Kimball
  - “Čišćenje podataka je problem broj jedan u skladištenju podataka”—DCI survey
- Osnovni zadaci
  - Popunjavanje nedostajućih vrijednosti
  - Identificiranje podatak izvan granica (outliers) i izglađivanje (smooth) šuma u podacima
  - Korekcija nekonzistentnih podataka
  - Rješavanje reudancije izazvane integriranjem podataka

Nedostajući podaci

- Podaci nisu uvijek raspoloživi
  - Npr., mnoge n-torke nemaju zabilježenu vrijednost za neke atribute (null vrijednost)
- Podaci mogu nedostajati zbog
  - Greške u radu opreme
  - Nekonzistentni s drugim podacima i zato izbrisani
  - Podaci nisu unešeni zbog nesporazuma
  - Određeni podaci nisu smatrani značajnim u trenutku unosa
  - Ne bilježi se povijet ili izmjene nad podacima
- Nedostajući podaci trebaju biti popunjeni.

Što uraditi s nedostajućim podacima?

- Ignorirati n-torku: obično se radi kada nedostaje oznaka klase (kod klasifikacije – nije djelotvorno ako postotak nedostajućih vrijednosti po atributu značajno varira).
- Ručno pounjavanje nedostajućih vrijednosti: zamorno + neizvedivo?
- Automatsko popunjavanje s
  - Globalnom konstantom: npr., “nepoznato”, nova klasa?!
  - Aritmetičkom sredinom vrijednosti atributa
  - Aritmetička sredina za sve uzorke koji pripadaju istoj klasi (mudrije)
  - Najvjerojatnija vrijednost: temeljeno na Bayesian formuli ili stablu odlučivanja

Šumovi u podacima

- Šum: random greška ili varijanca u varijabli koja se mjeri
- Netočna vrijednost atributa zbog
  - Greške na instrumentima za prikupljanje podataka
  - Problemi na unosu podataka
  - Problemi pri prijenosu podataka
  - Tehnološki limiti
  - Nekonzistentnost u imenovanju
- Drugi problemi koji zahtjevaju čišćenje podataka
  - Dupli slogovi
  - Nepotpuni podaci
  - Nekonzistentni podaci

Što uraditi sa šumovima u podacima?

- **Grupiranje (Binning)**
  - Prvo sortirati podatke i podijeliti u grupe (jednake frekvencije)
  - Nakon toga se može raditi izglađivanje (smooth) po aritmetičkoj sredini grupe, po medijani, po graničnim vrijednostim, itd.
- **Regresija**
  - Izglađivanje podešavanjem podataka regresijskim funkcija,a
- **Klasteriranje**
  - Otkrivanje i uklanjanje vrijednosti izvan granica (outliers)
- **Kombinirati računalnu i ljudsku provjeru**
  - Otkrivanje sumnjivih vrijednosti i provjera od strane ljudi (npr., rad s mogućim outliers)

Jednostavne metode diskretizacije: Binning

- **Jednaka širina (udaljenost) particioniranje**
  - Podijeliti raspon na *N* intervala jednake veličine: uniformni raster (grid)
  - Ako su *A* i *B* najniža i najveća vrijednost atributa, širina intervala će biti:

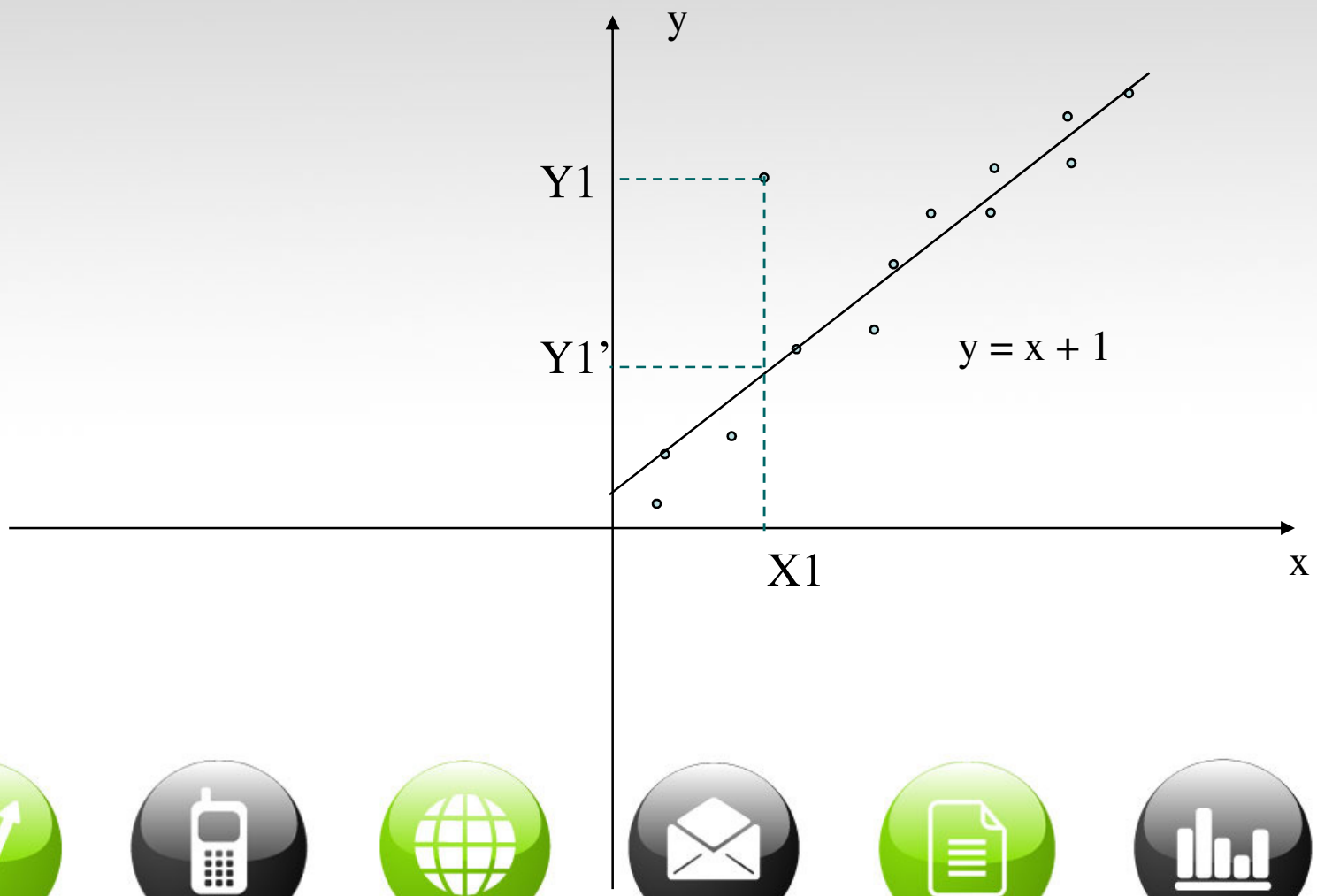
    $W = (B - A)/N.$
  - Najjasnija, ali outliers mogu dominirati
  - Nije preporučljivo za asimetrične podatke
- **Jednaka dubina (frekvencija) particioniranje**
  - Podijeliti raspon na *N* intervala, svaki sadrži približno isti broj uzoraka
  - Dobro skaliranje podataka
  - Rad s kvalitativnim podacima može biti zahtjevan

Metode grupiranja (binning) za izglađivanje (smoothing) podataka

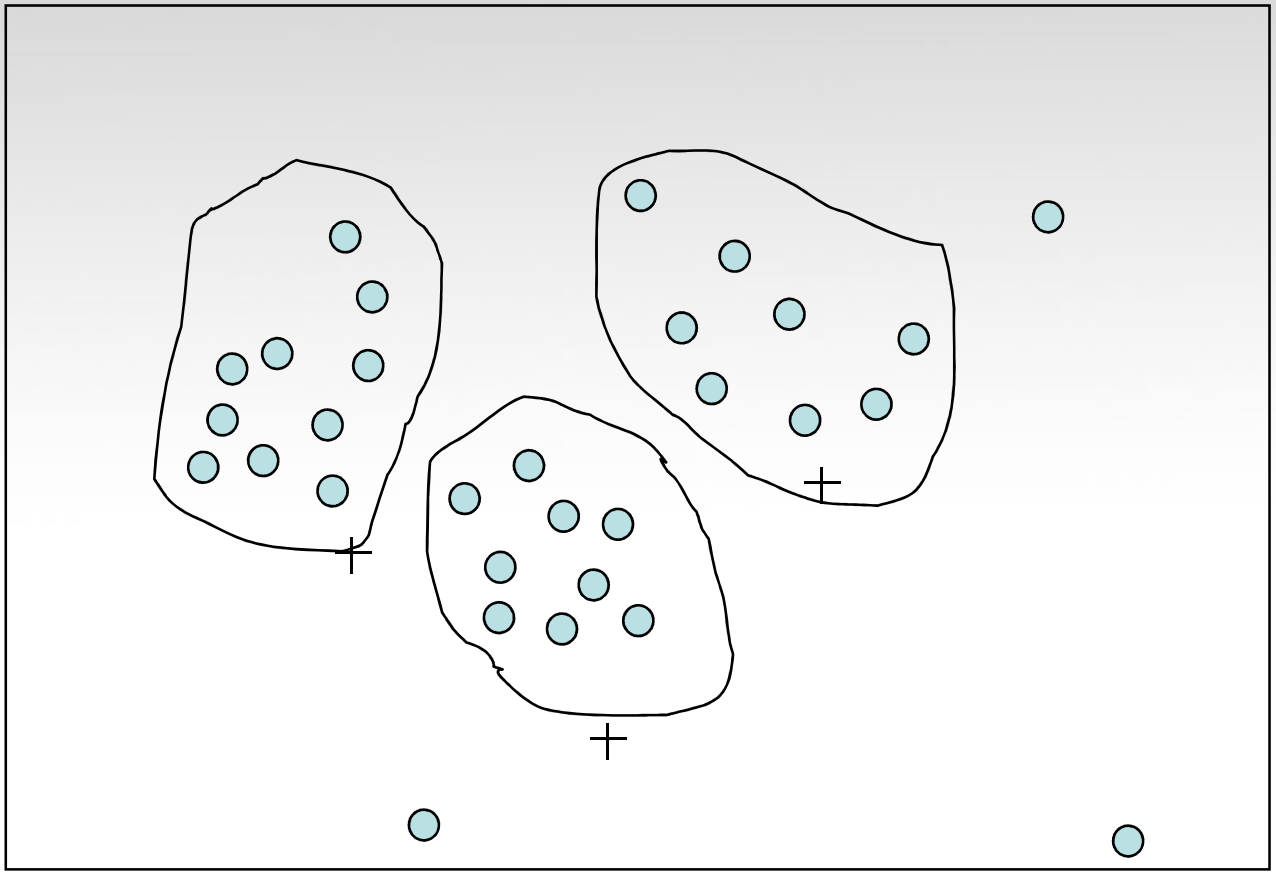
- Sortiranje podataka za cijene : 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Particioniranje u grupe jednake frekvencije:
  - Grupa 1: 4, 8, 9, 15
  - Grupa 2: 21, 21, 24, 25
  - Grupa 3: 26, 28, 29, 34
* Izglađivanje pomoću aritmetičke sredine grupe:
  - Grupa 1: 9, 9, 9, 9
  - Grupa 2: 23, 23, 23, 23
  - Grupa 3: 29, 29, 29, 29
* Izglađivanje pomoću krajnjih granica:
  - Grupa 1: 4, 4, 4, 15
  - Grupa 2: 21, 21, 25, 25
  - Grupa 3: 26, 26, 26, 34

Regresija



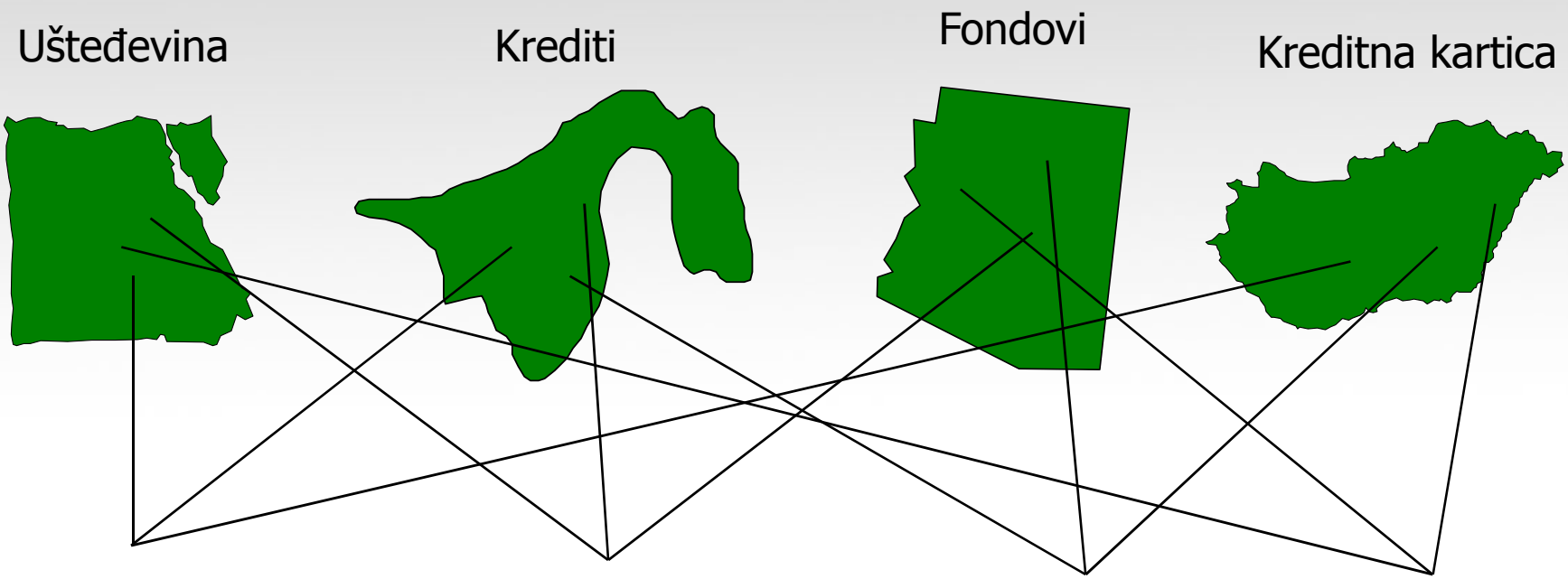
Klaster analiza



Integriranje podataka

- **Integriranje podataka:**
  - Kombiniranje podataka iz više izvora u jedno koherentno skladište
- **Integriranje shema: npr., A.cust-id ≡ B.cust-#**
  - Integriranje meta podataka iz različitih izvora
- **Problem identificiranja entiteta:**
  - Identificiranje stvarnih entiteta iz više izvora podataka, npr. Bill Clinton = William Clinton
- **Otkrivanje i rješavanje konflikata u podacima**
  - Za isti stvarni entitet, vrijednosti atributa iz različitih izvora su različite
  - Mogući razlozi: različito predstavljanje, različite skale, npr. metrička vs Britanska

Integracija podataka iz različitih izvora



Isti podatak
Različit naziv

Različit podatak
Isti naziv

Podaci na ovdje i
Nigdje drugo

Različiti ključevi
Isti podatak

Rješavanje redundancije pri integriranju podataka

- **Redundantni podaci se pojavlju vrlo često pri integriranju većeg broja baza podataka**
  - *Identificiranje objekta*: Isti atributi ili objekt mogu imati različite nazive u različitim bazama podataka
  - *Izvedivost podataka:* Jedan atribut može biti “derivirani” atribut u drugoj tablici, npr. godišnji prihod
- **Redundant ni atributi se mogu otkriti pomoću korelacije**
- **Pažljivo integriranje podataka iz više izvora može pomoći u smanjenju/izbjegavanju redundancije i nekonzistentnosti, te poboljšati brzinu i kvalitetu analize podataka**.

Transformiranje podataka

- Iglađivannje (Smoothing): uklanja šumove u podacima
- Agregacija: zbrajanje, izrada kocki
- Generalizacija: koncept penjanja po hijerarhiji
- Normalizacija: skaliranje “pada” unutar malog, definiranog ranga
  - min-max normalizacija
  - z-score normalizacija
  - Normalizacija pomoću decimalnog skaliranja
- Izrada atributa
  - Novi atributi nastaju na temelju zadanih

Transformiranje podataka: Normalizacija

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

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

  - Pr. Neka je prihod na rasponu 12,000 - 98,000 normaliziran na [0.0, 1.0]. Onda se 73,000 mapira u $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

- Z-score normalizacija (μ: mean, σ: standard deviation):

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

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

- Normalizacija pomoću decimalnog skaliranja

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

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)

Pojmovi transformiranja podataka

- 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 …

Pojmovi transformiranja podataka

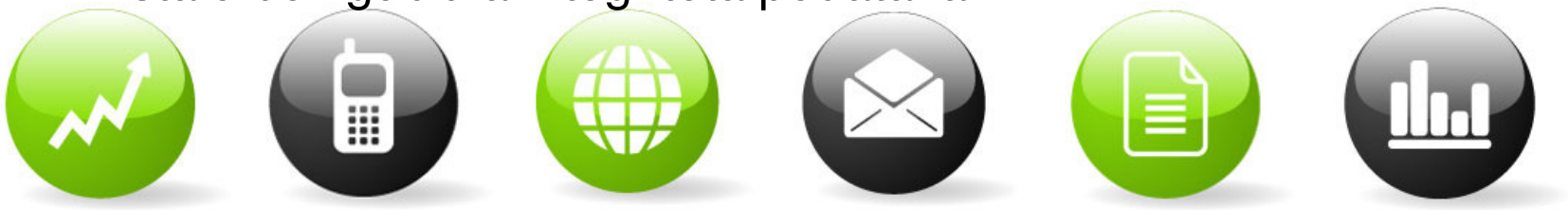
- Obogaćivanje
  - Korištenje podataka iz vanjskih izvora kako bi se obogatili operativni podaci.

- 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

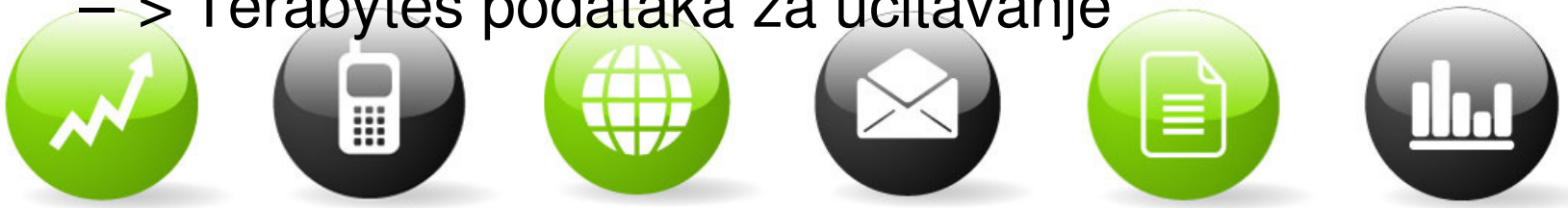
- Potpuno izdvajanje iz osnovnih tablica
  - Čita čitavu izvornu tablicu: preskupo
  - Možda jedini izbor za nasljeđene sustave

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: 21.11.2016.

1. Tema: Dimenzijsko modeliranje– priprema za diskusiju
2. Pripremiti prezentaciju svog projekta u trajanju od 5 min max:

   - use case dijagram

   - dijagram aktivnosti

PITANJA