

SKLADIŠTE PODATAKA - UVOD

Definicije

Skladište podataka (data warehouse): Domenski orijentisan, integrisan, vremensko promenljiv i neuništiv skup podataka namenjen podršci odlučivanju kod upravljanja nekim sistemom.

Podskladište podataka (data mart): Poseban izdvojeni deo skladišta podataka namenjen potrebama nekog dela sistema.

Web skladište podataka (data webhouse): Distribuirano skladište podataka implementirano preko web-a (za koje ne postoji centralizovano čuvanje podataka).

U sva tri slučaja, reč je o strateškom IS, za razliku od operacionog IS.

Pogodnosti

Visok stepen povraćaja ulaganja
Povećanje konkurentnosti
Povećanje produktivnosti odlučivanja
Povećanje kvaliteta odlučivanja

Poređenje

Operacioni (OLTP) IS

Trenutni podaci stanja i prometa
Detaljne podatke
Dinamički podaci
Ponavljajuća predefinisana obrada
Visok nivo transakcione aktivnosti
Predvidiv način korišćenja
Transakciono orijentisan
Aplikativno orijentisan
Podrška svakodnevnom odlučivanju
Opislužuje veliki broj korisnika

Strateški (OLAP) IS

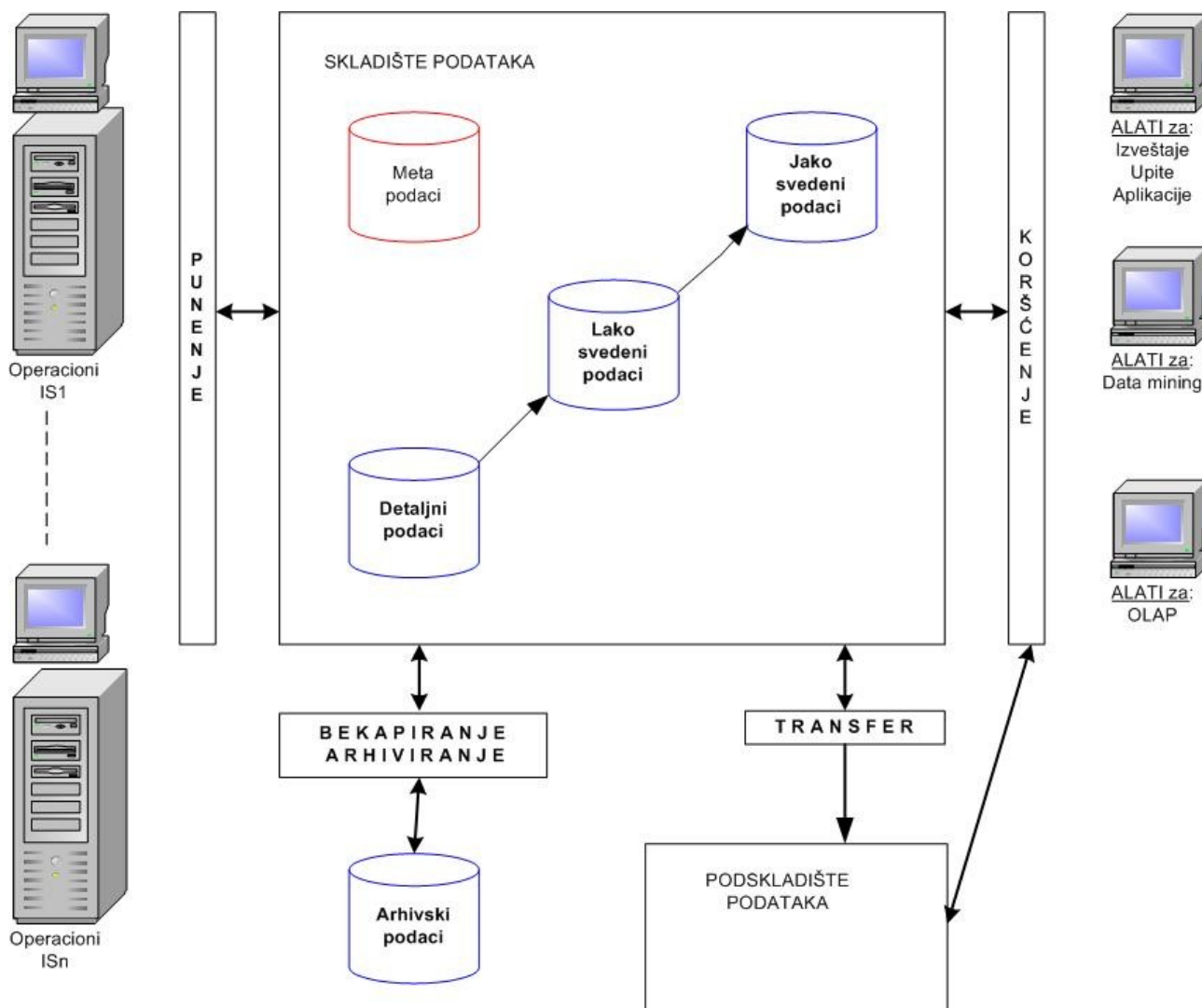
Istorijski podaci stanja i prometa
Detaljni i srednje i visoko svodni podaci
Ugalvnom statički podaci
Od hoc obrada po zahtevu
Nizak / srednji nivo transakcione aktivnosti
Nepredvidiv način korišćenja
Analitički orijentisan
Domenski orijentisan
Podrška strateškom odlučivanju
Opislužuje manji broj korisnika

Problemi koji se javljaju u vezi skladišta podataka

Podcenjivanje resursa potrebnih za punjenje podacima
Skriveni problemi unutar izvornih IS
Neobuhvatanje neophodnih podataka unutar izvornih IS
Semantika i homogenizacija podataka
Visoki zahtevi za resursima
Vlasništvo/pristup podacima
Obimno naknadno održavanje
Dugoročnost projekta (≥ 3 godine)
Kompleksnost integracije sistema

SKLADIŠTE PODATAKA – USTROJSTVO I RAD

Struktura skladišta podataka



DATA WAREHOUSE - skladište podataka (globalna namena)
DATA MART - podskladište podataka (specifična namena)

Implementacija skladišta podataka

Uobičajeno je da se vremenom iz jednog skladišta podataka formiraju podskladišta podataka, pri čemu isti podaci i dalje ostaju u skladištu podataka.

U praksi se dešava obrnuto - prvo se realizuju pojedina kritična podskladišta podataka koja se direktno pune iz operacionih IS, a zatim se naknadno formira skladište podataka. Od tog trenutka, tok podataka iz operacionih IS je ka skladištu podataka, a podskladišta podataka se pune transferom iz skladišta podataka.

Implementacija skladišta podataka je preko baze podataka (uglavnom relacione) sa visokim stepenom redundanse, a mehanizam svođenja je zasnovan na okidačima.

Punjenje skladišta podataka

U principu postoje dva načina punjenja podacima iz operacionih IS:

- ✓ **Totalno punjenje:** U određenim vremenskim trenucima, skladište se isprazni a zatim ponovo napuni podacima iz operacionih IS.
- ✓ **Inkrementalno punjenje:** Prilikom punjenja, u skladište se prenose samo izmene nastale u operacionim IS nakon prethodnog punjenja.

Postoje dve varijante inkrementalnog punjenja:

- ✓ **Paketno inkrementalno punjenje:** Vršiti se u određenim vremenskim trenucima. Zahteva izmene u operacionom IS (bazi podataka) koje će implementirati mehanizam prepoznavanja nastalih izmena.
- ✓ **Neprekidno inkrementalno punjenje:** Vršiti se neprekidno. Nakon svake promene u operacionim IS mehanizmom okidača vrši se prenos podataka ka skladištu podataka.

Mogući problemi kod punjenja skladišta podataka, naročito izraženi kod heterogenih operacionih IS, su:

- ✓ **Netačnost podataka iz operacionih IS:** Pri punjenju je neophodno filtriranje podataka, odnosno odbacivanje netačnih podataka.
- ✓ **Neusaglašenost podataka po tipu/preciznosti:** Pri punjenju je neophodno usaglašavanje po oba osnova.

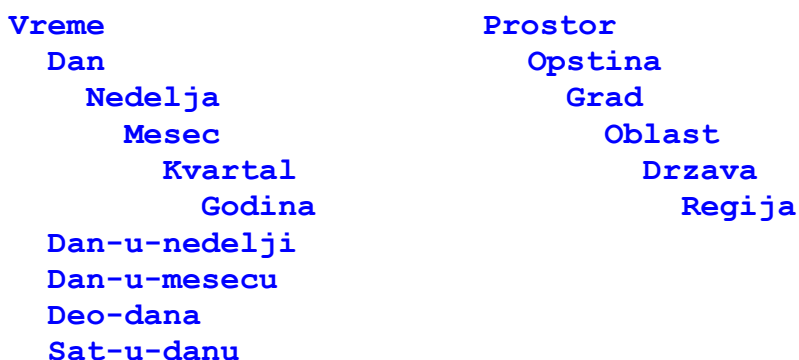
SKLADIŠTE PODATAKA – DIMENZIONIRANI MODELI

Dimenziono modeliranje

Polazi se od toga da se atributi entiteta (tabela) mogu podeliti na dve grupe:

- ✓ **atributi mere:** atributi čija vrednost odražava meru (veličinu) nešega; bitna osobina je mogućnost svođenja; mogu biti i izvedeni (primer: datum rođenja > starost)
- ✓ **atributi dimenzije:** atributi čija vrednost služi kao osnov klasifikacije i svođenja; mogu biti bez i sa varijabilnom granulacijom (nivoima svođenja).

Dva specijalna atributa dimenzije sa varijabilnom granulacijom: **Prostorni** i **Vremenski**



Prostorni dimenzioni atribut ima jednu šemu hijerarhije. Najniži nivo rezolucije je GPS geografska pozicija iz koje se dalje izvode viši nivoi rezolucije. U tom pogledu očekuje se uvođenje novog standardnog tipa podatka (POSITION?) i odgovarajućih standardnih funkcija izvođenja.

Vremenski dimenzioni atribut ima 5 paralelnih šema hijerarhije. Najniži nivo rezolucije je vreme (sat-minut-sekunda-...) i za to postoji standardni tip podatka TIMESTAMP. Postoji i čitav niz standardnih funkcija koje vrše izvođenje po raznim šemama hijerarhija (TIME, DATE, DAY_OF_WEEK, MONTH itd).

Kod dimenzionog modeliranja može istovremeno da postoji više vremenskih dimenzionih atributa (na primer, Dan-u-nedelji i Sat), pod uslovom da su one nezavisne, odnosno da nizu izvodive jedna iz druge (kombinacija Dan i Dan-u-mesecu nema smisla).

Dve vrste tabela u dimenzionom modelu:

- ✓ **tabela fakata** (FT, jedna): sastoji se iz složenog primarnog ključa u koji ulaze ID-ovi svih dimenzija i jednog ili više atributa mere;
- ✓ **tabele dimenzija** (DTi, više): sastoje se iz prostog primarnog ključa koji odgovara jednoj komponenti primarnog ključa tabele fakata i jednog ili više atributa dimenzije.

Unutar jednog skladišta ili podskladišta podataka može se nalaziti više dimenzionalnih modela.

Varijante dimenzionog modeliranja

Šema "zvezda" (star):

Sadrži jednu FT i za svaku dimenziju tačno jednu DT. DT mogu da sadrže denormalizovane podatke (neključne funkcijske zavisnosti).

FT --> DTi

Šema "pahuljica" (snowflake):

Sadrži jednu FT i niz DT koje ne mogu da sadrže denormalizovane podatke, odnosno reference na dodatne informacione tabele IT.

FT --> DTi --> ITj ...

Šema "zvezda-pahuljica" (starflake):

Kombinacija navedene dve varijante.

Sazvežđe (constellation):

Situacija kada kod više dimenzionalnih modela različite FT referišu iste DT.

SKLADIŠTE PODATAKA – DIMENZIONA ANALIZA (OLAP)

Osnovu OLAP dimenzione analize predstavlja OLAP kocka (CUBE) koja može biti sa 1 ili n dimezija (vizualizacija je moguća do n=3).

City	Time	Total Revenue
Glasgow	Q1	29726
Glasgow	Q2	30443
Glasgow	Q3	30582
Glasgow	Q4	31390
London	Q1	43555
London	Q2	48244
London	Q3	56222
London	Q4	45632
Aberdeen	Q1	53210
Aberdeen	Q2	34567
Aberdeen	Q3	45677
Aberdeen	Q4	50056
.....		
.....		

(a)

		City			
Quarter	City	Glasgow	London	Aberdeen	
	Q1	29726	43555	53210	
	Q2	30443	48244	34567	
	Q3	30582	56222	45677	
	Q4	31390	45632	50056	

(b)

Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
.....			
.....			

(c)

		City			
Property Type	City	Glasgow	London	Aberdeen	
	Flat	15056	14555	14578	15890
	House	14670	15888	16004	15500
					
					

(d)

Primer: dvodimenziona i trodimenziona OLAP kocka za promet nekretnina.