Data warehouse koncepti

1. DEFINICIJA I POJAŠNJENJE DWH
2. OLTP VS OLAP
3. DATA WAREHOUSE MOGUĆNOSTI
4. OLAP KOCKE, DIMENZIJE, MERE, TIPOVI (ROLAP, MOLAP, HOLAP))
5. DWH ŠEMA TIPOVI (STAR, SNOWFLAKE)
6. INMNON/KIMBALL PRISTUP

Predavači: Manja Miljević – Data Engineer

Oktobar 2023.

1) Definicija I pojašnjenje DWH

Glavni uzroci pojave analitičke baze podataka su:

- potrebe za pretvaranjem transakcionih podataka u informacije,
- omogućava veću inicijativu krajnjim korisnicima, koji ne moraju da znaju SQL i detalje o postavci informacionog sistema,
- zahtevi za automatizacijom rada analitičara, komercijalista i menadžera.

Analitičke baze podataka uslovile su nastanak skladišta podataka (Data Warehouse – DWH), sa glavnim ciljem izveštavanje korišćenjem transakcionih podataka, istorijskih podataka, kao i podataka iz drugih sistema. Podaci u analitičkoj bazi podataka su **denormalizovani.** Na osnovu redundantnih podataka dolazi se do bržeg pretraživanja podataka, tj. odgovori na upite su brži u odnosu na relacionu bazu podataka. DWH ne služi za onlajn transakcije u smislu ažuriranja (izmena), unosa novih slogova ili brisanja postojećih. Ta skladišta predstavljaju neku vrstu prečišćenih podataka prilagodjenih izvršenju u cilju izrade što većeg broja razlicitih izveštaja. Sistemi koji obezbeđuju izveštaje koji su rezultat sinteze velikog broja podataka, obično su podrška nekom strateškom odlučivanju, nazivaju se OLAP sistemi kojima su izvor podataka upravo skladišta podataka (DWH). Od analitičke baze podataka se očekuje da obezbedi odgovore na brojne zahteve korisnika. Kako bi se to omogućilo, uvodi se **multidimenzionalni pristup podacima.** Prilikom objašnjavanja dimenzionalnog modela često se pribegava modelu “kocke” i na taj način se stiče utisak o velikom broju dimenzija.

2) OLTP vs OLAP

Karakteristika	Operativna BP (OLTP)	OLAP
Tipične operacije	Dnevne operacije, unos, brisanje, ažuriranje	Kompleksni upit
Kritične transakcije	Operativna obrada	Informativna obrada
Ažuriranje BP	Veliki broj DML operacija (Insert, Delete, Update)	Punjenje i periodično osvežavanje “istorijski”, sumarizirani, multidimenzionaln, integrisani, konsolidovani
Frekvencija upita	Niska / srednja	Visoka
Kompleksnost upita	Niska	Visoka
Obim baze podataka	Mala / srednja (100 MB – GB)	Velika (100GB – TB)
Očekivano vreme odgovora	Mili sekunde - sekunda	Nekoliko sekundi – više sati
Vremenska diskretizacija podataka	Dan - sekunda	Dan - godina
Aktuelnost podataka	Do jedne godine	Više godina
Pristup / broj korisnika	Čitanje-pisanje / hiljade	Čitanje / stotine
Granularnost podataka	Elementarni podaci	Agregirani podaci
Šema BP	Normalizovana, kompleksnija,relaciona	Denormalizovana, manje kompleksna
Izvori podataka	Tekuće poslovanje	Operativna BP, interni i eksterni izvori
Podrška poslovnim procesima	Operativno poslovanje	Analiza i odlučivanje

3) *Data Warehouse mogućnosti*

Skladište podataka (DWH) je prvenstveno fokusirano na to da vrhovni menadžment pravovremeno efektivno i efikasno donosi strategijske odluke.

Osnovne prednosti:

a) *donosi raznovrsnost i bogastvo informacija,*
b) *ušteda vremena,*
c) *obezbeđuje konzistentnost i bolji kvalitet podataka,*
d) *obezbeđuje „istorijske“ informacije*
e) *generiše visok ROI (profitabilnosti investicije)*

Dodatne prednosti DW sistema predstavljaju:

a) agregirani podaci,
b) evaluativni podaci,
c) tvrdi i meki podaci.

4) OLAP kocke, Dimenzije, Mere, Tipovi (ROLAP, MOLAP, HOLAP)

Osnovna ideja dimenzionalnog modelovanja je da gotovo svaki tip poslovnih podataka može da bude predstavljen u vidu kocke u kojoj ćelije kocke sadrže vrednosti a ivice kocke predstavljaju dimenzije podataka.

Multidimenzionalni pogled pruža analitičarima gledanje podataka iz hijerarhijske perspektive. Na taj način se omogućava segmentiranje baze podataka i to: određivanje podskupova na osnovu zadatih kriterijuma („dicing“), rotaciju („data slicing“), kretanje kroz hijerarhiju („drill-up“, drill-down) i drugo.
Osnovni pojmovi koji su vezani za dimenzionalno modelovanje su:

1) Činjenice – uglavnom predstavljaju numeričke mere poslovanja, ali postoje izuzeci kada mogu da budu i kvalitativne, tekstualne;
2) Dimenzije – predstavljaju deskriptivne informacije u vidu teksta o svakom redu u tabeli činjenica. Neke od opštih dimenzija su: vreme, roba, partner, tržišni segment i tome slično;
3) Relacije – na osnovu njih zaključuje se da su tabele činjenica u interakciji sa tabelama dimenzija;
4) Granularnost – daje uvid na kom nivou detaljnosti su uskladištene činjenice u skladištu podataka.

1. ROLAP

Relaciona online analitička obrada ROLAP alati pristupaju podacima u relacionoj bazi podataka i generišu SQL upite da bi obradili informacije na željenom nivou, onda kada je to korisniku potrebno. Kod njih se izostavljaju faze prethodne obrade i dodatnog memorisanja podataka kao u drugim varijantama (npr. MOLAP). Sa ROLAP alatima mogu se kreirati dodatne tabele u bazi podataka (najčešće sumarne tabele odn. agregacije) koje sumarizuju podatke za bilo koju željenu kombinaciju dimenzija.

Prednosti ROLAP su:

A. pogodan je za obradu velikih količina podataka, posebno kada postoji nadređeni entitet u bazi (parent table) koji ima mnogo elemenata podređenih entiteta (child table);
B. u poređenju sa drugim OLAP mogućnostima, vreme ekstrahovanja podataka je značajno kraće, iz razloga što postoje veliki izbor alata za ekstrahovanje podataka i mogućnosti preciznog prilagođavanja ETL koda konkretnom modelu podataka;
C. podaci su u standardnoj relacionoj bazi podataka pa alat za pristup ne mora biti OLAP alat, može im se pristupiti bilo kojim SQL alatom za izveštanjave;
D. pogodan za neagregirane podatke;
E. razdvajanjem memorijskog prostora za podatke od multidimenzionalnog modela mogu se uspešno modelovati podaci koji se inače teško uklapaju u striktno dimenzioni model;

F. ROLAP može da pojača kontrolu autorizacije baze podataka kao što je sigurnost nivoa reda tabela, uz filtriranje rezultata upita prema unapred postavljenim kriterijuma (SQL uslov WHERE).

Nedostaci su:

- ✓ generalno, obrada podataka je sporija nego kod drugih varijanti;
- ✓ potrebno je dodatni napor za razvoj ETL koda zbog formiranja agregiranih tabela;
- ✓ vreme odziva na upit raste ukoliko se izostavi korak kreiranja agragiranih tabela, iz razloga što se onda moraju čitati velike tabele detaljnih;
- ✓ kako su SQL alati oslonjeni na SQL za sve obrade, oni nisu podesni onda kad ima dosta kalkulacija koje se teže izvode u SQL-u, npr. budžetiranje, alokacija, finasijsko izveštavanje, obračun amortizacije i sl.

Proizvodi koji su zasnovani na ROLAP:

- Microsoft Analysis Services,
- MicroStrategy,
- Oracle Business Intelligence Suite Enterprise Edition (ranije Siebel Analytics)
- Mondrian (besplatan ROLAP server).

2. MOLAP

MOLAP takođe podržava ROLAP multidimenzionalni model podataka. Glavna razlika u odnosu na ROLAP tehnike je u tome što MOLAP zahteva prethodnu obradu i memorisanje podataka u vidu OLAP kocke. Većina MOLAP proizvoda podatke smešta u memoriju u vidu optimizovanog multidimenzionalnog niza, a ne u relacionu bazu podataka.

Prednosti MOLAP su:

A. odziv na upite sporo traje, zahvaljujući optimizovanoj memoriji, multidimenzionalnom indeksiranju i keširanju;
B. uz pomoć tehnika kompresovanja potreban je manji prostor na disku u odnosu na podatke koji su u relacionoj bazi podataka manji;
C. automatsko izračunavanje višeg nivoa agregacija podataka;
D. pogodan je da obimne podataka svede na podatke manjih dimenzija;
E. modeli nizova obezbeđuju prirodno indeksiranje;
F. efikasna ekstrakcija podataka kroz prestrukturisanje agregiranih podataka.

Nedostaci su:

- u nekim MOLAP rešenjima korak učitavanja podataka, u okviru ETL procesa može dugo da traje, pogotovo kada je u pitanju velika količina podataka.

- MOLAP alati pokazuju pad performansi upita na modelima podataka sa visokom kardinalnošću
- redundantni (ponavljajući) podaci mogu da se jave u nekim MOLAP

Proizvodi bazirani na MOLAP:

- Cognos Powerplay,
- Oracle Database OLAP Option,
- Microsoft Analysis Services (Excel, SSRS)
- Razni alati za isvestavanje (Sharepoint, Pyramid, Quick Sense, Quick View…)

3. HOLAP

HOLAP alati mogu pristupati i relacionim i multidimenzionim bazama podataka. Cilj korišćenja HOLAP alata jeste da se iskoriste prednosti MOLAP alata (kratko vreme odziva sistema i analitičke mogućnosti) i ROLAP alata (dinamički pristup podacima). Sve u svemu, to je ROLAP koji ima mogućnost izvršavanja vrlo složenih SQL naredbi. Glavni zadatak je bio da se doda mogućnost za rad sa analitičkim bazama podataka, a ujedno da se zadrže i prednosti ROLAP-a.

Proizvodi bazirani na HOLAP:

- Microsoft Analysis Services,
- Oracle Database OLAP Option,
- MicroStrategy
- Microsoft SQL Server 7.0 OLAP Services (podržava hibridni OLAP server).

	Skladištenje detaljnih podataka	*Skladištenje sumarnih/ agregiranih podataka*	*Potreban skladišni prostor*	*Vreme odziva na upit*	*Brzina procesiranja*	*Latentnost*
MOLAP	Multidimenzionalne BP	Multidimenzionalne BP	Srednji (detaljni podaci se kompresuju pre skladištenja	Kratko	Brzo	Visoka
HOLAP	Relacione BP	Multidimenzionalne BP	Mali	Srednje	Brzo	Srednja
ROLAP	Relacione BP	Multidimenzionalne BP	Veliki	Dugo	Sporo	Niska

5) DWH šema tipovi (Star, Snowflake)

Šema zvezde i šema pahulje predstavljaju dve osnovne šeme u dimenzionalnom modelovanju. Šema zvezde predstavlja osnovnu arhitekturu dimenzionalnog modelovanja. Zvezdasta šema ima strukturu koja je dosta jednostavna, sa veoma malim brojem tabela. Dimenzionalna šema je vrlo razumljiva, kako za analitičare i same korisnike.

I) Dimenzionalni model zvezdaste šeme sadrži dve vrste tabela: **tabele činjenice (FACT table) i tabele dimenzija (Dimension Table)**. Šema zvezde je dizajnirana na takav način, da se u sredini nalazi tabela činjenica koja je okružena tabelama dimenzija. Svaka **tabela dimenzija** ima jedinstveni primarni ključ koji se tačno podudara sa jednim od delova složenog ključa u tabeli činjenica. **Tabela činjenica** sadrži u zavisnosti od potreba i zahteva klijenta detaljne ili agregirane, konzistentne podatke. Podaci mogu biti normalizovani ili denormalizovani. Ukoliko su normalizovani tada tabela činjenica sadrži strane/spuštene ključeve (Foreign Key)

Delovi tabele činjenica su:

a) Numeričke činjenice, koje mogu biti aktuelne i izvedene;
b) Spoljni/spušteni ključevi, koji povezuju svaku pojedinačnu činjenicu sa njenom pripadajućom vrednosti u tabeli dimenzija. U tabeli dimenzija predstavlja primarni ključ (Primary key);
c) Degenerativni ključ, npr. broj naloga;

Tabele dimenzija su manjeg obima i sadrže podatke o jednoj od dimenzija kocke. One su opisnog karaktera (špifarnici). Komponente tabele dimenzija su:

a) Primarni ključ, najčešće je to promenljiva inkrementalnog tipa;
b) Vrednost produkcionog ključa – koji se upotrebljava za referenciranje;
c) Opisne atribute;
d) Hijerarhijske atribute;
e) Zaglavlja reda i ograničenja.

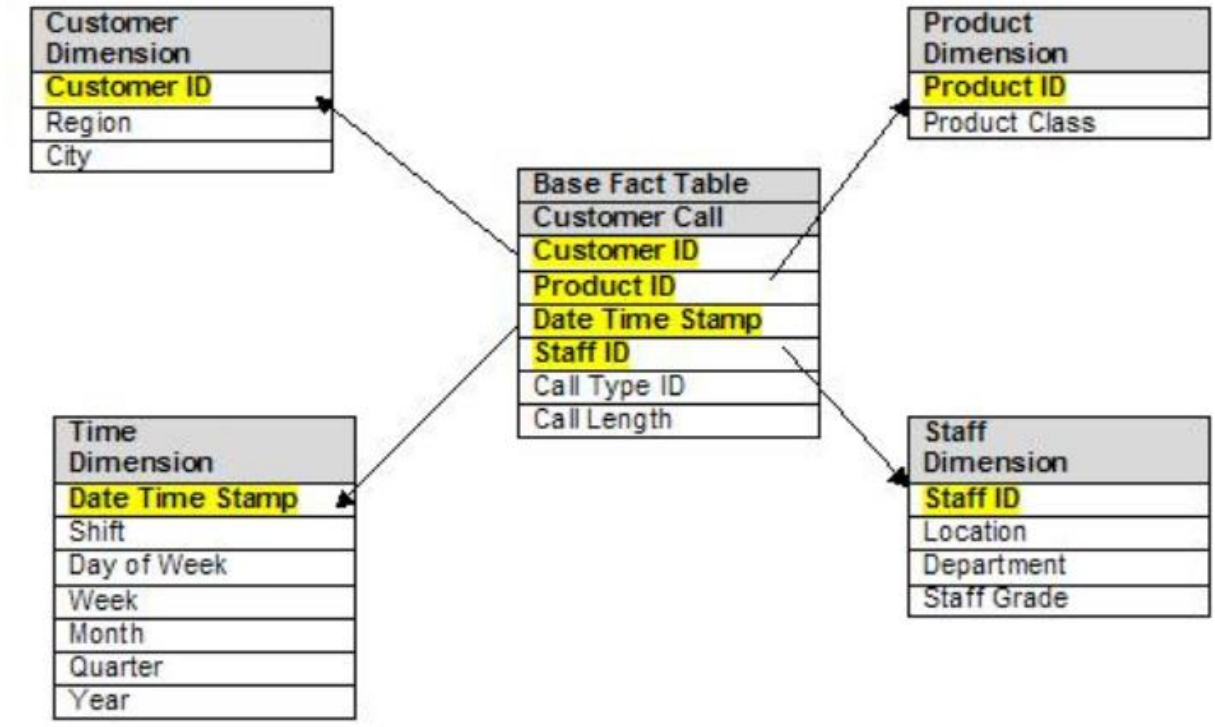
Prednosti:

- Omogućava da veoma jednostavan model podataka definiše kompleksnu multudimenzionalnu strukturu podataka;
- Zvezdastom shemom se lako definišu hijerarhijski odnosi unutar svake dimenzije, a to pojednostavljuje kreiranje veza kroz više tabela;
- Pored toga, zvezdasta shema smanjuje broj fizičkih veza u procesu upita poboljšavajući time performansu;
- U pojednostavljenom modelu podataka korisnik lakše formuliše upite pomoću kojih će dobiti potrebu informaciju;
- Zvezdasta shema omogućava da se Data Warehouse širi uz relativno malo održavanje.

Nedostaci:

a) Ne predstavlja univerzalno rešenje za dizajn DW baze podataka;
b) Nekada ne predstavlja najbolju tehniku dizajniranja baze podataka;
c) Nije pogodno rešenje kada tabele dimenzija sadrže veliki broj redova i atributa;

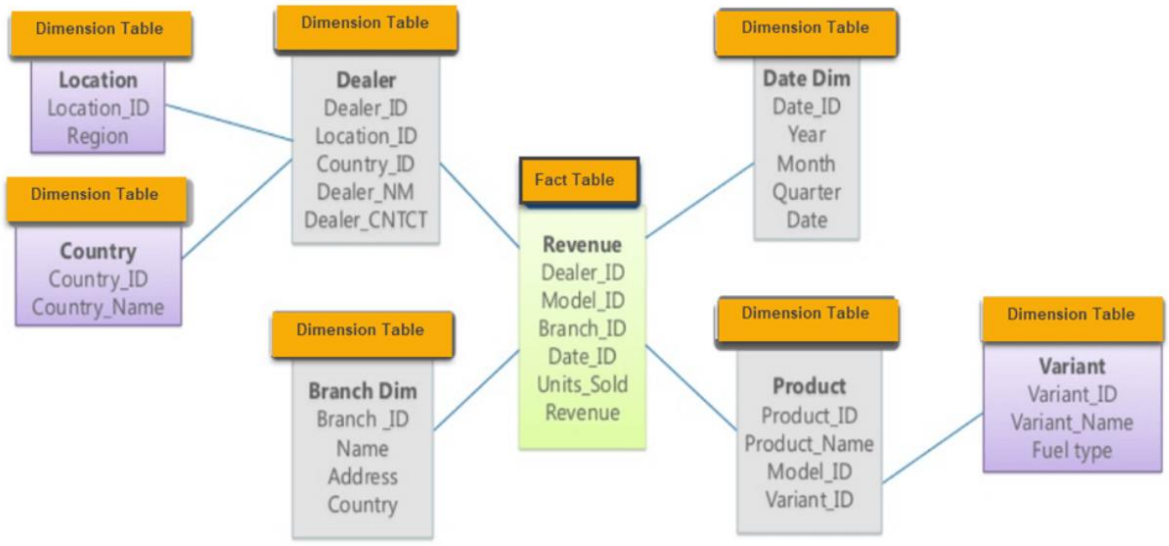
d) S obzirom da su podaci denormalizovani, ponavljanje grupa podataka za veliki broj atributa dovodi do ogromnog proširenja tabele činjenica.



II) Šema pahulje predstavlja varijantu zvezdaste šeme u kojoj su i tabele činjenica i tabele dimenzija u trećoj normalnoj normalnoj formi, odnosno potpuno su normalizovane, ali sama struktura tabele činjenica ostaje ista. Pahuljasta šema nastaje tako što se polja koja se nalaze u nekim tabelama dimenzije dele na posebne tabele. Ovim se postiže viši nivo normalizacije, ali se ujedno i smanjuje performantnost a neretko i jednostavnost korišćenja alata. Postoje i multiple tabele činjenica koje karakteriše postojanje više tabela činjenica koje su međusobno povezane preko tabela dimenzija.

Dimenzionalni model je proširljiv za uključivanje neočekivanih novih elemenata podataka i nove odluke o dizajnu:

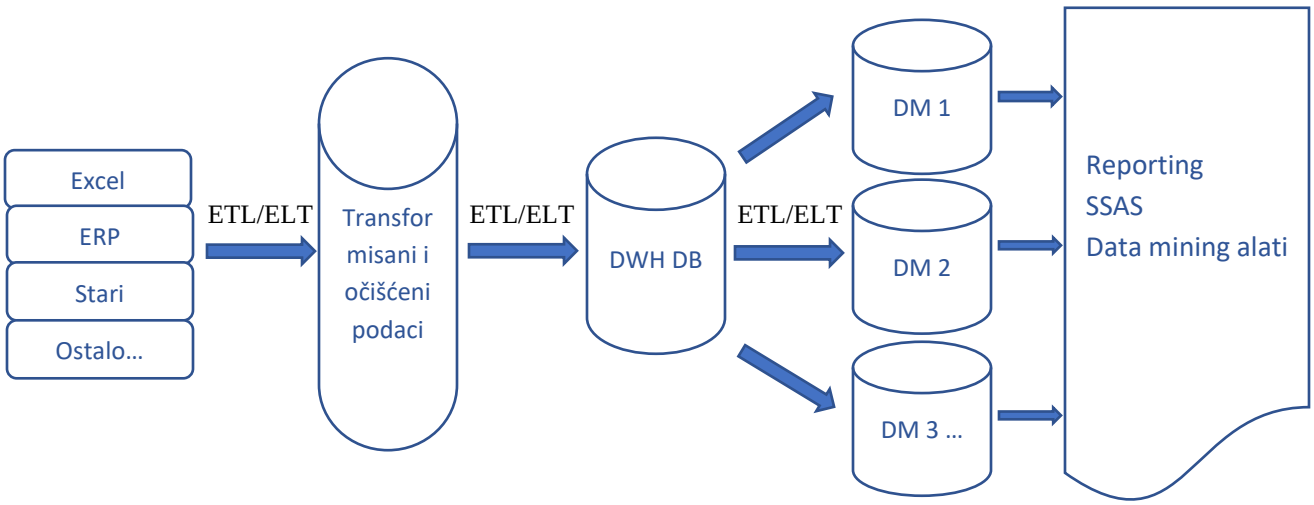
- dodavanje novih neočekivanih podataka ako su konzistentni sa granulacijom postojeće tabele činjenice;
- dodavanje potpuno nove dimenzije ako je jedinstvena vrednost te dimenzije definisana za svaki postojeći zapis činjenica;
- dodavanje novih dimenzionalnih atributa;
- razbijanje postojećih zapisa dimenzije na niže nivoe granulacije od nekog vremena unapred.


Dimension Table
Location
Location_ID
Region
Dimension Table
Country
Country_ID
Country_Name
Dimension Table
Dealer
Dealer_ID
Location_ID
Country_ID
Dealer_NM
Dealer_CNTCT
Dimension Table
Branch Dim
Branch _ID
Name
Address
Country
Fact Table
Revenue
Dealer_ID
Model_ID
Branch_ID
Date_ID
Units_Sold
Revenue
Dimension Table
Date Dim
Date_ID
Year
Month
Quarter
Date
Dimension Table
Product
Product_ID
Product_Name
Model_ID
Variant_ID
Dimension Table
Variant
Variant_ID
Variant_Name
Fuel type

6) *Inmnon/Kimball pristup*

A) ***Inmnon (CIF arhitektura),*** kreira centralni Data Storage (DWH) pa iz toga proizilaze DM (Data Martove sastoji iz nekoliko komponenti, a to su:

- Podaci operativnog Sistema
- Zona pripreme podataka
- Centralno skladište podataka (DWH)
- Ispostave podataka (DM)
- Alati za pristup podacima



Prvi nivo:

- Podaci iz opertaivnog sistema – obično su to transakcioni podaci, relacione baze podataka, podaci se skupljaju iz više različitih operativnih sistema (starih, novih…), excel, csv fajlovi…

Drugi nivo:

- Sprovodi se transformacija i čišćenje podataka, postiže se njihova konzistentnost.
- Nad podacima se sprovodi skup pravila da bi se dobili u željenom formatu koji je pogodan kasnije za korišćenje od strane poslovnih korisnika. Takvi podaci se čuvaju u posednom skladištu.
- Nad ovim podacima se ne sprovode DML procedure.
- Nad ovim spremištem podataka ne postoje servisi za postavljanje upita
- To su obično flat fajlovi i relaciona baza podataka

Treći nivo:

- Iz okruženja gde su smešteni transformisani i očišćeni podaci, podaci se prebacuju u skladište podataka (DWH), koje predstavlja izvor podataka za različite ispostave, u ovom slučaju u DM – Data Mart-ove.

- S obzirom da skladište (DWH) predstavlja spremište podataka na nivou čitave kompanije, ono mora sadržati atomske podatke u trećoj normalnoj formi (3NF).

Četvrti nivo:

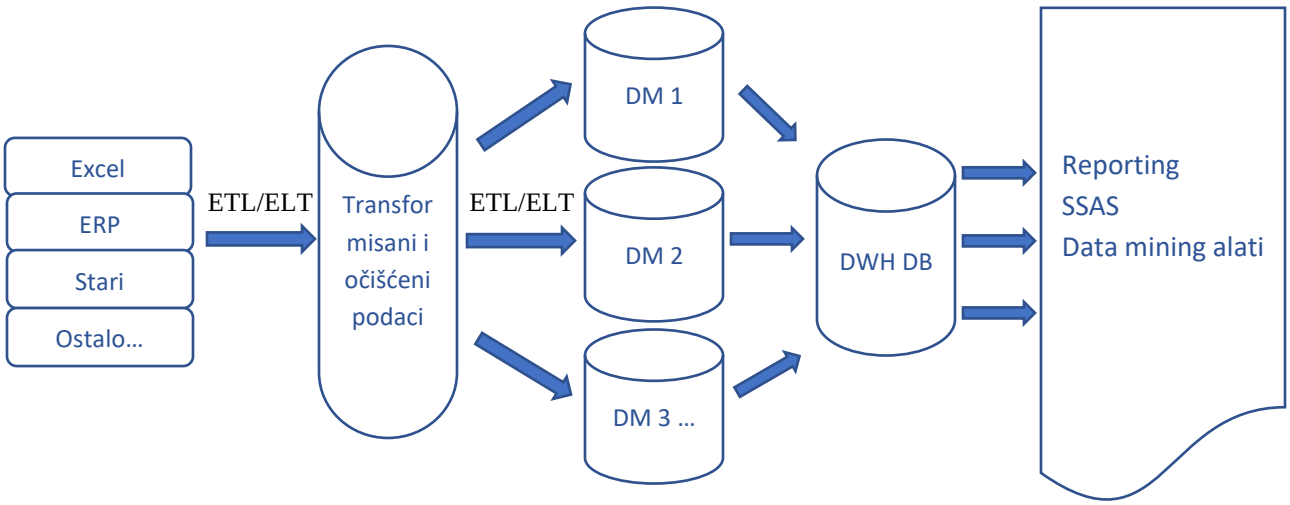
- Ispostave podataka (DM) treba, sa druge strane, da služe kao spremište podataka koje zadovoljava potrebe određenih organizacionih jedinica i podaci u njima poseduju višedimenzionalnu strukturu, u vidu zvezdastih i pahuljičastih šema.

Peti nivo:

- Podaci koji su smešteni u ispostave podataka mogu se koristiti za razna analitička izveštavanja, predviđanja, upotrebom široke palate alata predviđenih za to

B) ***Po Kimball-u (BUS arhitektura)*** se, okruženje skladišta podataka (Data Warehouse Environment) sastoji iz nekoliko komponenti, a to su:

- Izvorni operacionalni sistemi,
- Zona pripreme podatka (Transformisani i očišćeni podaci),
- Zona prezentacije podataka podeljeno po poslovnim procesima (Data Mart),
- Centralnoo skladišete podataka
- Alati za pristup podacima (Data Access Tools).



Prvi nivo:

- Podaci iz opertaivnog sistema – obično su to transakcioni podaci, relacione baze podataka, podaci se skupljaju iz više različitih operativnih sistema (starih, novih…), excel, csv fajlovi…

Drugi nivo:

- Zona pripreme podataka ne podrazumeva samo zonu skladištenja podataka, već ovde spadaju i tzv. ETL (Extract/Transfer/Load) procesi tj. procesi ekstrahovanja, transformisanja i učitavanja podataka. Ekstrahovanje podrazumeva isčitavanje i razumevanje izvornih podataka i njihovo kopiranje u zonu pripreme, kako bi se njima dalje manipulisalo.
- Nakon ekstrahovanja vrši se transformisanje podataka koje podrazumeva čišćenje podataka, kombinovanje podataka iz različitih izvora, brisanje duplikata, generisanje ključeva unutar skladišta i slično. Pomenute transformacije prethode učitavanju podataka u zonu prezentacije.
- Zona pripreme podataka se nalazi između operacionalnih sistema i zone prezentacije, i korisnici nemaju mogućnost da joj pristupaju.
- Nad ovim spremištem podataka ne postoje servisi za postavljanje upita
- To su obično flat fajlovi i relaciona baza podataka

Treći nivo:

- Iz okruženja gde su smešteni transformisani i očišćeni podaci, podaci se prebacuju u skladište podataka (DWH), koje predstavlja izvor podataka za različite ispostave, u ovom slučaju u DM – Data Mart-ove.
- Zona pripreme podataka se nalazi između operacionalnih sistema i zone prezentacije, i korisnici nemaju mogućnost da joj pristupaju. Podaci u njoj mogu biti u trećoj normalnoj formi (koju propagira Inmon), mada se Kimball protivi ovakvom stavu smatrajući da se na ovaj način stvari komplikuju i da se u ovom slučaju ETL procesi bespotrebno izvršavaju dvaput: prvi put prilikom učitavanja u normalizovanu bazu, a zatim ponovo kada je potrebno učitati dimenzionalne modele (prezentacija zahteva postojanje dimenzionalnih struktura).
- Kimball insistira da se pri tome koristi višedimenzionalni model podataka (i to striktno zvezdasta šema) i takođe ističe da ispostave podataka ne smeju da sadrže samo agregirane podatke, već moraju sadržati i atomske podatke.

Četvrti nivo:

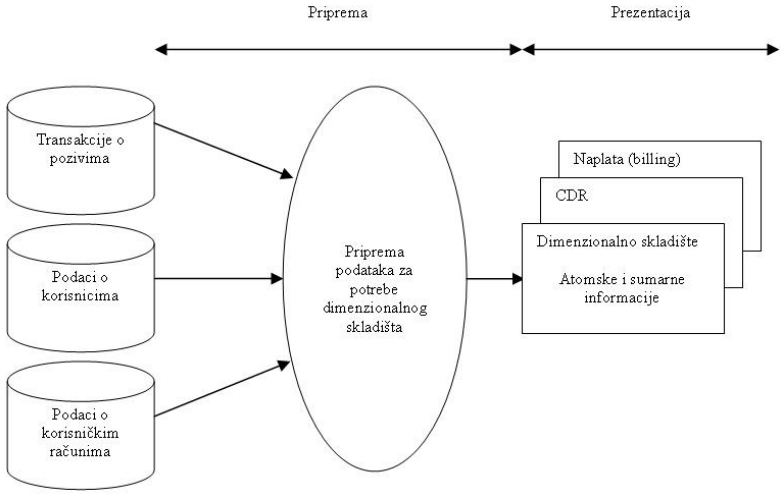
- Ispostave podataka (DM) treba, sa druge strane, da služe kao spremište podataka koje zadovoljava potrebe određenih organizacionih jedinica i podaci u njima poseduju višedimenzionalnu strukturu, u vidu zvezdastih i pahuljičastih šema.

Peti nivo:

- Podaci koji su smešteni u ispostave podataka mogu se koristiti za razna analitička izveštavanja, predviđanja, upotrebom široke palate alata predviđenih za to
- Alati za kreiranje Ad-hoc upita
- Modeliranje, predviđanje, vredbovanje, data mining

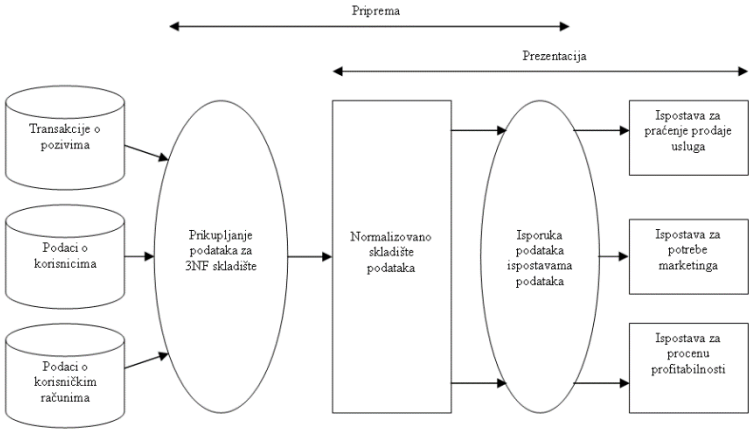
Komparacija

- Bitno je da se u skladištu čuvaju samo oni podaci koji su potrebni donosiocima odluka. Obe arhitekture takođe polaze od toga da je jako važno prvo definisati opštu arhitekturu skladišta (na nivou kompanije), iako će se ono implementirati po fazama.
- Kod Kimballa podaci se transformišu u informacije u sloju pripreme podataka, pri čemu podaci potiču iz različitih operacionalnih izvora.



Dimenzionalno skladište podataka (BUS arhitektura)

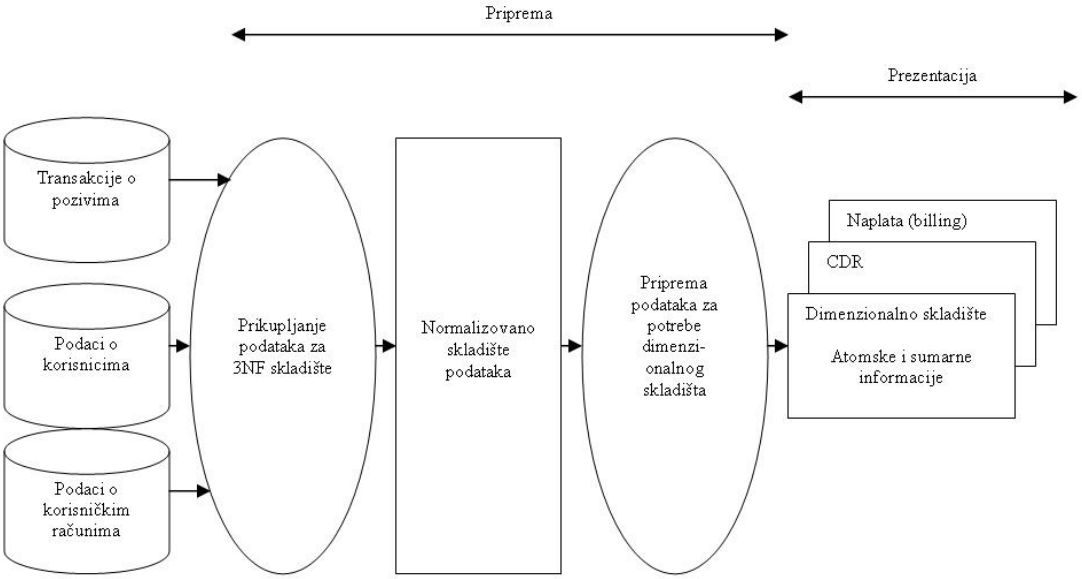
- Skladište podataka ima dimenzionalnu strukturu. Ovakav dimenzionalni model sadrži iste informacije kao normalizovani model, ali su korisnički paketi implementirani tako da se mogu jednostavno koristiti i upiti lako izvršavati.
- Dimenzionalni model pored atomskih informacija uključuje i sumarne informacije. Upiti idu do najnižeg nivoa detaljnosti
- Kod Inmnona CIF arhitektura koristi iste izvore podataka kao BUS, međutim podaci se prevode u treću normalnu formu (3NF) i ti se atomski podaci uvode u skladište.
- Normalizovano skladište obezbeđuje podatke za ispostave podataka. Ispostave zadovoljavaju potrebe pojedinih poslovnih funkcija (odnosno odeljenja) unutar kompanije, i one praktično određuju koje će se sumarne informacije posmatrati.
- Dakle, atomski podaci imaju sasvim drugačiju strukturu u odnosu na sumarne, što nije bio slučaj kod BUS arhitekture.



Normalizovano skladište podataka sa dimenzionalnim ispostavama koje sadrže sumarne informacije (CIF arhitektura)

	Inmnon – *CIF arhitektura*	**Kimball – *BUS arhitektura***
Učitavnje dimenzionalnog modela	Normalizacija podataka	Dimenzionalna struktura
Atomistika podataka	Atomski podaci su uskladišteni u normalizovani DWH	Atomski podaci moraju imati dimenzionalnu strukturu
DWH	Normalizovani	Konformisane dimenzije
Prednosti	Čvrsta osnova za BI, doslednost podataka na svim poljima podataka	Brži uvid, olakšana analitika, izveštavanje za pojedinačne poslovne procese/timove
Nedostaci	Visoki početni troškovi i značajno vreme za izgradnju	Mogućnost suvišnih podataka I nedostatak doslednosti podataka pošto se razvijaju autonomno.

- Moguće je primeniti i hibridni pristup, tj. usvojiti najbolje karakteristike obe arhitekture. Iz CIF-a se preuzima normalizovano skladište podataka, ali mu pridodaje dimenzionalno skladište koje je prisutno kod BUS arhitekture, koje će skladištiti atomske i sumarne podatke. Osnovni nedostatak ovog pristupa leži u redudantnosti atomskih podataka. Zbog toga ga uglavnom koriste kompanije koje su već implementirale normalizovano skladište podataka, ali žele i da koriste dobre strane BUS-a.



Hibridni pristup