NAČRTOVANJE VEČDIMENZIONALNIH PODATKOVNIH BAZ

Sebastian Lahajnar, Alenka Rožanec
PRIS Inženiring, Ljubljana, Upravna enota Ljubljana
E-pošta: sebastian.lahajnar@pris-inz.si, alenka.rozanec@gov.si

Izvleček

Dandanes obstajajo različne tehnike za modeliranje podatkov, ki se večinoma nanašajo na relacijske in objektne podatkovne baze. Poleg njih pa vse bolj pridobivajo na pomenu predvsem večdimenzionalne baze, kot pomemben del sistemov OLAP. Namen članka je predstavitev osnovnih konceptov večdimenzionalnih podatkovnih baz, katerih dobro razumevanje je ključ do uspešnega modeliranja in implementacije. Članek obravnava tudi osnovne korake večdimenzionalnega modeliranja, od razumevanja obstoječe situacije in določitve ciljev, preko definiranja dimenzij, hierarhij in članov, do končne definicije kompleksnih formul.

Abstract

Today there are many different techniques for data modeling, which are mostly related to relational and object databases. There is also another type of databases in expansive growth, named multidimensional databases, as an important part of OLAP systems. The purpose of the paper is to present all the basic concepts of multidimensional databases, good understanding of which is critical for successful modeling and implementation. The paper also covers basic steps for multidimensional modeling from understanding the current situation and determining goals, through defining dimensions, hierarchies and members, to the final definition of complex formulas.

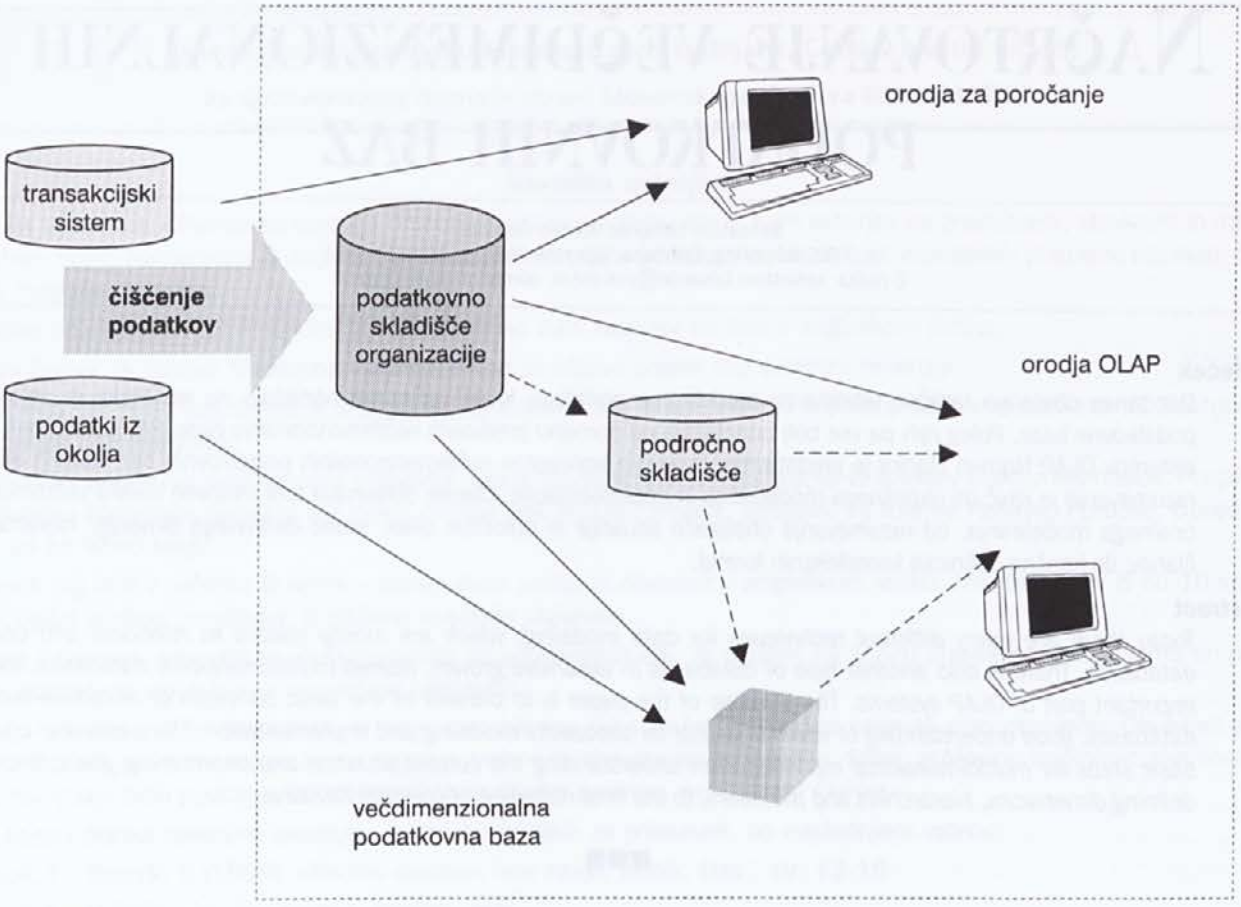
1. Uvod

Potrebe po pravočasnih in kvalitetnih informacijah pri spremljanju poslovanja, pripravi planov za bodoče poslovanje ter sprejemanju poslovnih odločitev so v preteklosti pripeljale do množične uporabe različnih programskih orodij, predvsem preglednic in raznovrstnih poročil na osnovi podatkov iz relacijskih podatkovnih baz. Vendar pa ne ene ne druge ne vsebujejo določenih funkcionalnosti, potrebnih za tovrstne analize. Zato so bile že npr. preglednicam dodane vrtilne tabele, ki omogočajo večdimenzionalne poglede ter obračanje podatkov na različne načine, z namenom pridobiti čimveč koristnih informacij. Naslednja stopnja tega razvoja pa so vsekakor orodja OLAP (On Line Analytical Processing), ki skušajo končnim uporabnikom zagotoviti kar najkvalitetnejše informacije iz množice nakopičenih podatkov transakcijskih sistemov in podatkovnih skladišč.

2. Direktorski informacijski sistemi

Direktorski informacijski sistemi (slika 1), katerim so orodja OLAP v prvi vrsti tudi namenjena, dandanes temeljijo na podatkovnih skladiščih. Potreba po izgradnji podatkovnih skladišč izhaja iz številnih pomanjkljivosti klasičnih transakcijskih sistemov, ki so namenjeni predvsem zbiranju podatkov, ne pa njihovi analizi. V transakcijskih sistemih imamo opravka z veliko količino podrobnih podatkov, pri čemer je vedno vprašljiva njihova točnost in medsebojna konsistentnost. V večjih sistemih (več sto tabel) z več tisoč transakcijami dnevno lahko obseg podatkovne strukture in količina podatkov hitro presežeta obvladljive meje, rezultat česar so zapletene, posledično pa tudi počasne poizvedbe. V nasprotju s transakcijskimi sistemi hranijo podatkovna skladišča (ponavadi v obliki relacijske podatkovne baze) velike količine prečiščenih in agregiranih podatkov ter so na ta način primeren podatkovni vir tako za klasična poročila, kot tudi za zapletene analize preteklega poslovanja in napovedovanje prihodnosti.

Znotraj direktorskega informacijskega sistema najdemo različne tipe podatkovnih tokov, pri čemer vsi potekajo v smeri od transakcijskega sistema in zunanjega okolja h končnemu uporabniku. Dandanes imajo v večini organizacij še vedno največjo vlogo orodja za poročanje, ki omogočajo izdelavo različnih poročil neposredno nad transakcijskimi podatki ali nad podatkovnim skladiščem, seveda kadar obstaja. Orodja OLAP so po drugi strani najbolj učinkovita, če imajo podatke shranjene v večdimenzionalnih podatkovnih bazah ali posebej prirejenih področnih skladiščih podatkov (data marts), ki so namenjena le posameznim delom poslovnega sistema.



Slika 1: Shema direktorskega informacijskega sistema

Orodja OLAP večinoma povezujemo z večdimenzionalnimi podatkovnimi bazami, na načrtovanje katerih se bomo osredotočili v nadaljevanju, pri čemer je potrebno poudariti, da je OLAP koncept uporabniškega vmesnika, ne pa načina shranjevanja podatkov. Tako glede na podporno tehnologijo ločimo dva tipa orodij OLAP:

- **ROLAP** (Relational OLAP): podatke črpajo neposredno iz relacijske podatkovne baze, v ozadju analiz, pregledovanj se nahaja poizvedovalni jezik SQL. Hitrost poizvedb je odvisna od postavljenih indeksov, nivoja predhodnega združevanja in je v povprečju slabša kot pri orodjih MOLAP, saj SQL ni optimiziran za izvajanje tipičnih operacij OLAP, kot so vrtanje v globino, rotacija itd.
- **MOLAP** (Multidimensional OLAP): podatki so shranjeni v večdimenzionalni podatkovni bazi, ki temelji na strukturi večdimenzionalnih polj, posebej prirejeni za izvajanje analiz OLAP. Primerna so za hranjenje manjših količin podatkov (omejeno število celic), poizvedbe so konstantno hitre (velika stopnja predhodnega združevanja podatkov), ažuriranje baze je počasnejše, saj je potrebno ponovno preračunavanje vseh agregatov.

Zastavlja se vprašanje, kateri pristop hranjenja podatkov je boljši. Seveda je odgovor odvisen od številnih faktorjev. V primeru izgradnje vseobsegajočega podatkovnega skladišča je to vsekakor relacijska baza podatkov, če pa gre za podporo le določenega dela poslovnega sistema, pa je primernejša uporaba večdimenzionalne baze. Tehnologiji v bistvu nista tekmeci, pač pa se medsebojno dopolnjujeta, kar dokazujejo tudi sami proizvajalci orodij OLAP, ki ponujajo izdelke v obeh kategorijah (podjetje Oracle ima na primer ROLAP orodje Discoverer in MOLAP orodje Express). Izkušnje zadnjih let so pokazale, da je izgradnja obsežnih, centraliziranih podatkovnih skladišč zelo zahtevna in dolgotrajna, kar pogosto privede do neuspeha celotnega projekta. Iz omenjenega razloga se podjetja danes

večinoma usmerjajo v izgradnjo omejenih podatkovnih skladišč za posamezne dele poslovnega sistema, pri čemer pa imajo orodja MOLAP prednost.

3. Značilnosti orodij OLAP

Da neko programsko orodje uvrstimo v kategorijo orodij OLAP, mora vsebovati predvsem naslednje funkcionalnosti:

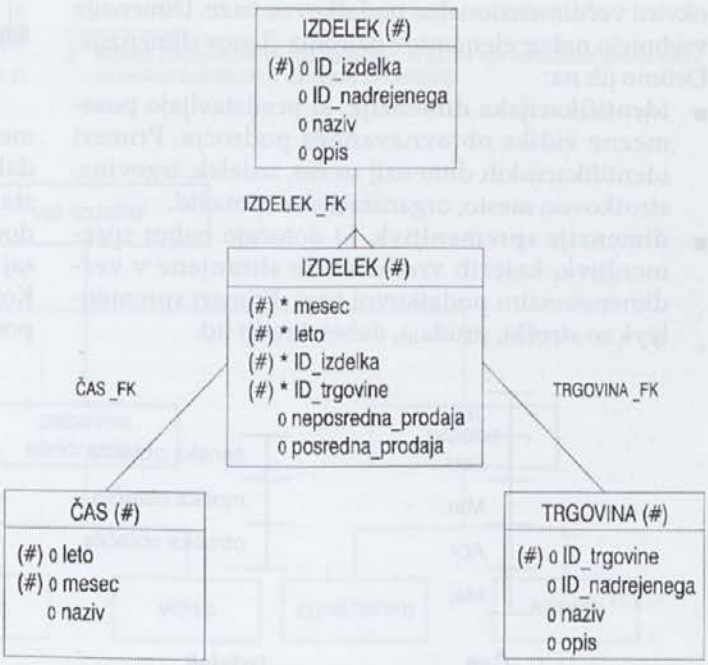
1. **Večdimenzionalnost:** omogoča uporabnikom pregled vrednosti različnih kazalnikov poslovne uspešnosti podjetja (npr. doseganje prodajnih kvot), primerjave s podatki iz preteklosti ali napovedanimi podatki za prihodnost po posameznih dimenzijah in njihovih hierarhijah. Za analize OLAP je potrebno predhodno pripraviti ustrezno podatkovno strukturo, kamor se zapišejo agregirani podatki, ki so potem na voljo uporabnikom pri različnih analizah.
2. **Hiter dostop in kalkulacije ter vrtanje v globino:** Poizvedovanje z orodji OLAP poteka na enostaven način, brez pisanja zapletenih stavkov SQL, zato ga lahko izvaja vsakdo. Uporabnik prehaja med različnimi nivoji podatkov, od popolnih proti podrobnejšim podatkom, kar imenujemo vrtanje v globino ali angleško drill-down. Odzivni časi so pretežno konstantni ne glede na vrsto poizvedbe in ne smejo biti večji od nekaj sekund, saj morajo slediti miselnemu procesu uporabnika. Čim boljše odzivne čase se skuša doseči z različnimi metodami: večdimenzionalne baze, združene tabele, ustrezno indeksiranje in podobno.
3. **Močne analitične sposobnosti:** Orodje mora poleg osnovnih seštevanj in povprečij po hierarhijah dimenzij vsebovati tudi naprednejše funkcije (npr. statistične) za finančne, prodajne in druge analize.
4. **Prožnost:** Je še ena od ključnih komponent sistemov OLAP in vključuje: različne načine pregledovanja podatkov (v obliki matrik, različnih vrst grafov, tabel s poljubno razmestitvijo stolpcev in vrstic), rotacijo ali kockanje (podatkovna struktura ne pogojuje prikaza, kot je to značilno v primeru preglednic), prožnost definicij (enostavno definiranje formul, oblikovanje števil itd.), prožnost analiz ter prilagodljiv, intuitiven in uporabniško prijazen vmesnik.
5. **Večuporabniški dostop:** Večina današnjih sistemov OLAP je tipa odjemalec-strežnik, kar zagotavlja hkraten dostop do podatkov in njihovo obdelavo večjemu številu uporabnikov.

4. Zvezdna shema

Koncept zvezdne strukture podatkovnega skladišča kot osnove za zajem podatkov v večdimenzionalne podatkovne baze se je v preteklih letih izkazal kot ena najboljših možnih rešitev. Samo ime, zvezdna shema, izhaja iz oblike podatkovnega modela, ki je zgrajen iz obsežne osrednje tabele (tabela dejstev) obkrožene z večjim številom pomožnih tabel (dimenzijske tabele). Razmerje med vsako posamezno dimenzijsko tabelo in tabelo dejstev je 1:N, pri čemer razmerja med dimenzijskimi tabelami niso podana. Tak način modeliranja podatkovnih skladišč, imenovan tudi dimenzijsko modeliranje, pomeni dobro osnovo za izgradnjo večdimenzionalnih modelov, saj že vnaprej ponuja možen nabor dimenzij, atributov in hierarhij.

Zvezdna shema tako vedno vsebuje dva tipa tabel:

- **Tabela dejstev:** osrednja tabela sheme, ki vsebuje večinoma številčne vrednosti, na osnovi katerih se bodo izvajale analize. Atributi tabele se delijo na atribute, ki določajo ključ (ponavadi sestavljen iz ključev dimenzijskih tabel) in atribute, ki nosijo neki semantični pomen.
- **Dimenzijske tabele:** tabele, ki predstavljajo posamezne vidike sistema oziroma različne poglede, skozi katere se bo izvajala analiza. Razen ključa vsebujejo večinoma tekstovne podatke, ki podrobneje specificirajo posamezno razsežnost.



Slika 2: Primer preproste zvezdne sheme za trgovsko podjetje

Slika 2 prikazuje preprosto zvezdasto shemo za trgovsko podjetje. Tabela dejstev vsebuje podatke o neposredni oziroma posredni prodaji strankam, ki so lahko predmet nadaljnje analize, tri dimenzijske tabele (čas, izdelek in trgovina) pa nakazujejo, s katerih vidikov se bo analiza izvajala: s časovnega, glede na kraj prodaje in glede na tip izdelka. Iz sheme je možno nadalje razbrati, da so izdelki in trgovine hierarhično urejeni, saj vsak zapis vsebuje tudi identifikator nadrejenega.

5. Osnovni koncepti večdimenzionalnih podatkovnih baz

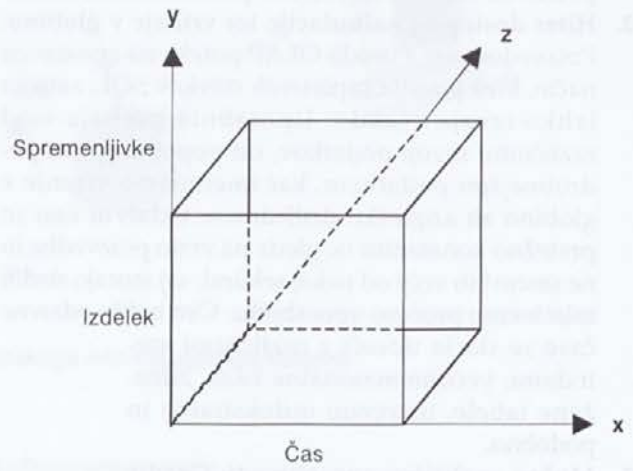
V nadaljevanju so predstavljeni osnovni koncepti večdimenzionalnih podatkovnih baz, katerih dobro razumevanje je ključnega pomena za izgradnjo učinkovitih rešitev MOLAP. Ker so večdimenzionalne podatkovne baze oziroma orodja tipa MOLAP podvrsta širšega pojma OLAP, so praktično vsi opisani koncepti uporabni tudi v primeru izgradnje rešitev, temelječih na relacijskih podatkovnih bazah (ROLAP). Razlike se pojavijo šele pri podrobni obravnavi posameznih konceptov (dimenzij, hierarhij itd.), saj različna tehnološka podlaga različno vpliva na funkcionalne značilnosti posameznih rešitev.

Dimenzije

Dimenzije so temelj tehnologije večdimenzionalnih podatkovnih baz. Primerna analogija zanje so atributi primarnega ključa v relacijski teoriji (enolično določajo posamezno n-terico znotraj relacije), saj tudi presek članov dimenzije enolično določa posamezno celico v okviru večdimenzionalne podatkovne baze. Dimenzije vsebujejo nabor elementov oziroma članov dimenzije. Delimo jih na:

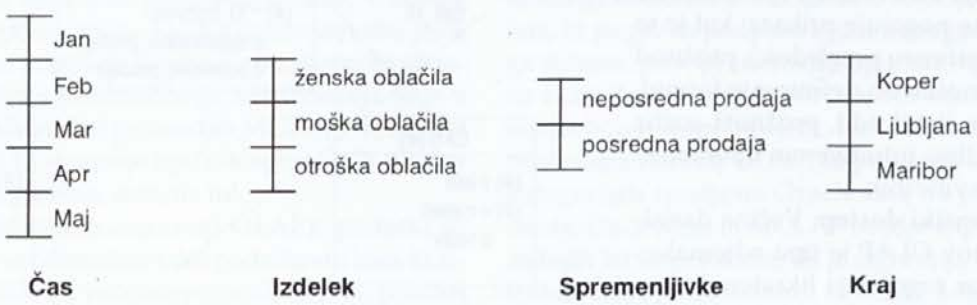
- **identifikacijske dimenzije**, ki predstavljajo posamezne vidike obravnavanega področja. Primeri identifikacijskih dimenzij so čas, izdelek, trgovina, stroškovno mesto, organizacijska enota itd.
- **dimenzije spremenljivk**, ki določajo nabor spremenljivk, katerih vrednosti so shranjene v večdimenzionalni podatkovni bazi. Primeri spremenljivk so stroški, prodaja, debet, kredit itd.

Pogosto kot sinonim za večdimenzionalno podatkovno bazo uporabljamo izraz hiperkocka – kocka z več kot tremi dimenzijami. Uporaba kocke (tri razsežnega koordinatnega sistema) kot prispodobe za predstavitev večdimenzionalnih podatkovnih baz je posledica človekovega naravnega razumevanja prostora, ki je omejeno na tri, v najboljšem primeru štiri razsežnosti, če za četrto razsežnost vzamemo čas. Predstavitveni problemi se pojavijo, ko želimo v poslovni model obravnavanega okolja vključiti četrti, peti itd. vidik, kar je v vsakdanji praksi prej pravilo kot izjema. Rešitev je v miselnem preskoku, kjer je potrebno geometrijsko predstavo dimenzije, ki se naslanja na fizične razsežnosti v smislu širine, višine in dolžine, zamenjati z logičnim konceptom dimenzij oziroma s sistemom večdimenzionalne strukture domen – MDS (Multidimensional Domain Structure).



Slika 3: Večdimenzionalna podatkovna baza predstavljena v obliki kocke

MDS omogoča prikazovanje poljubnega števila dimenzij. Vsaka izmed njih je predstavljena z navpično daljico, pri čemer je posamezen član dimenzije predstavljen z intervalom na daljici. Z uporabo MDS je dodajanje dimenzij ranga štiri in več povsem trivialno, saj preprosto dodamo novo dimenzijo v obliki daljice. Kombinacija intervalov vseh daljic na ta način določa posamezno celico večdimenzionalne podatkovne baze.



Slika 4: Primer predstavitve dimenzij z MDS

Hierarhije

Drugi temeljni koncept večdimenzionalnih podatkovnih baz so hierarhije. Le-te so osnova za združevanje podatkov in na ta način zagotavljajo možnost obravnavanja shranjenih podatkov na različnih kumulativnih ravneh. Čeprav definiranje hierarhij ni nujno potrebno v postopku izdelave večdimenzionalne baze, pa v vsakdanji praksi pogosto srečujemo hierarhično urejene dimenzije kot so čas, izdelki itd.

Hierarhično drevo vsebuje tri tipe elementov: koren, vozlišče in list. Listi drevesa (analitični člani dimenzije) se skladno s postavljenimi pogoji združujejo v vozlišča na različnih ravneh (sintetični člani dimenzije) vse do vrhnjega elementa – korena. Hierarhije so lahko simetrične ali asimetrične, pri čemer je v realnem svetu asimetričnih prav gotovo več. Posamezna dimenzija ima lahko eno ali več hierarhij, pri čemer vsaka združuje podatke po nekem drugem ključu. Tako imamo lahko na primer oblačila hierarhično urejena po spolu (moška, ženska, otroška), vrsti (hlače, puloverji, srajce, itd.), sezoni (zimska, letna) itd.

Opredeljevanje hierarhij je prav gotovo eno izmed najbolj zahtevnih opravil pri izgradnji večdimenzionalne podatkovne baze, saj s tem podrobneje določimo posamezne vidike (dimenzije) obravnavanega sistema. Izbor primernega števila ravni združevanja v marsičem vpliva na ustrezno predstavitev podatkov, ki je v direktorskih sistemih odločujoča za sprejemanje pravilnih poslovnih odločitev. Prav zato je ključnega pomena, da se v vseh vsebinskih podrobnostih posvetujemo z uporabniki, katerim bo sistem namenjen.

Hierarhije tudi pomembno vplivajo na strukturo, dosegljivost in smeri pregledovanja podatkov, ki je določena s številom neposrednih sosedov[1] posamezne celice. Če je za večdimenzionalno podatkovno bazo z n dimenzijami, pri čemer ima vsaka dimenzije le eno raven (brez hierarhije), značilno, da ima vsaka celica 2n sosedov, se število sosedov v primeru obstoja hierarhij poveča in se izračuna po formuli:

```
Št. sosedov = skupni produkt (število sosedov v posamezni
dimenziji + 1) – 1
```

Večje število sosedov pomeni večje število možnih smeri zaporednega pregledovanja podatkov, pri čemer pa obstaja tudi možnost neposrednega iskanja določene vsebine pod danimi pogoji. Posebna vrsta pregledovanja je vrtanje v globino, ki omogoča pregledovanje podatkov od kumulativ proti večjim podrobnostim.

Podatki

Večdimenzionalna podatkovna baza lahko vsebuje različne tipe podatkov (številski, tekstovni itd.), ki jih z vidika njihove pripadnosti delimo na:

- podatke, ki pripadajo posamezni celici in
- podatke, ki pripadajo članom dimenzij (atributi).

V prvem primeru gre za tipične, ponavadi številske podatke, ki predstavljajo osnovne ali izračunane vrednosti posameznih spremenljivk in s tem vrednostno opredeljujejo posamezen poslovni rezultat (na primer v trgovini A so v letu 1999 prodali 200 čevljev). Atributi po drugi strani identificirajo in podrobneje opisujejo posamezne lastnosti članov dimenzije in so zato večinoma tekstovni (na primer trgovina A je v lasti Matjaža Trgovca, ima 100 m^2 in je opredeljena kot trgovina z obutvijo).

1 *Sosedi posamezne celice so celice, ki so od nje oddaljene eno enoto v katerikoli kombinaciji dimenzij in hierarhij.*

Slika 5: Primer hierarhično urejene dimenzije

Formule

Celice v večdimenzionalni podatkovni bazi lahko vsebujejo dva tipa podatkov in sicer vhodne podatke (prenesejo se preko povezav iz posameznega podatkovnega izvora) ter izpeljane podatke, ki so rezultat preračunavanja vhodnih podatkov po izbranih formulah. Za razliko od klasičnih preglednic, kjer je formula dejansko lastnost posamezne celice (zapisana je v celici), formule v večdimenzionalnih bazah definiramo na nivoju dimenzij oziroma posameznih njenih članov. Tako neka formula ne pripada zgolj posamezni celici temveč vsem celicam, ki si delijo istega člana dimenzije. Formula posamezne celice je v tem primeru kombinacija vseh formul članov dimenzij, ki celico enolično opredeljujejo.

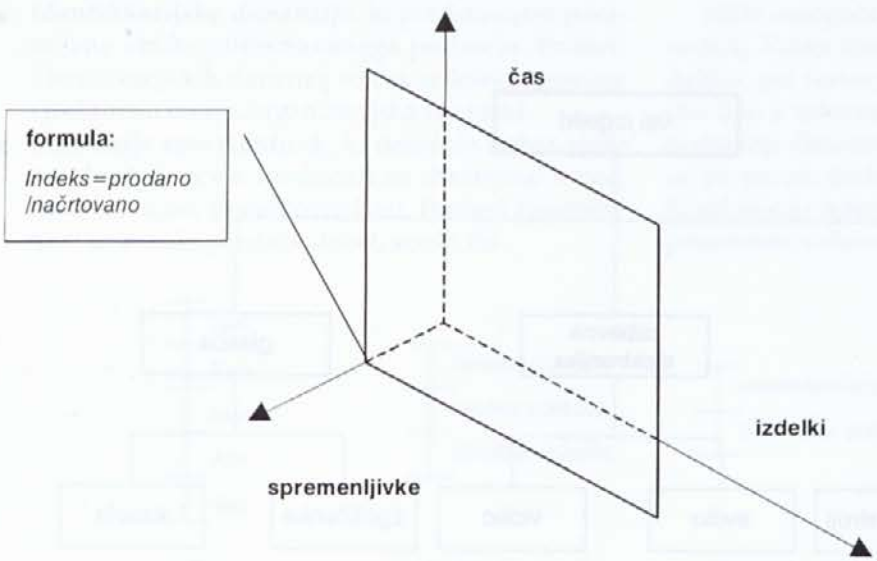
Problem, ki se lahko pojavi pri tovrstni obravnavi formul, je vrstni red njihovega izvajanja. Takoj ko imamo opravka z malo zahtevnejšimi formulami (kombinacija seštevanja, odštevanja, množenja in deljenja), je potrebno posebno pozornost posvetiti vrstnemu redu, saj so rezultati v posameznih primerih povsem različni. Primer prikazuje slika 7. V tabeli želimo izračunati indekse dejanske prodaje glede na načrtovano po posameznih izdelkih in skupno. V primeru, ko najprej seštevamo in nato delimo dobimo skupni rezultat 1, sicer pa 2. Brez doma je pravilen prvi rezultat, saj naša prodaja na žalost ni bila tako visoko nad načrtovano, kot kaže nepravilni izračun. Ne obstaja neko splošno pravilo, ki bi za vsak možen primer določalo ustrezen vrstni red. Potrebno je preprosto vsebinsko razmisliti o tem, kaj pravzaprav želimo izračunati in skladno s tem zagotoviti ustrezen vrstni red.

	prodano	načrtovano	indeks (P/N)
Hlače	120	100	1,2
Srajce	80	100	0,8
Skupaj	200	200	1 ali 2

Slika 7:
Neujemanje rezultatov v primeru različnega vrstnega reda izvajanja formul

Formule v povezavi s hierarhijami pa so tudi vzrok pojavu, ki ga poimenujemo eksplozija izpeljanih vrednosti. V tipičnih hierarhijah kot so čas, izdelki, kupci itd. je razmerje med vsemi člani dimenzije in člani, ki tvorijo liste drevesa (inflacijsko razmerje) med 1,5 in 2,5. (primer s slike 5: 10/6=1,67). V primeru večdimenzionalne podatkovne baze s tremi hierarhičnimi dimenzijami in povprečnim razmerjem 2 se razmerje poveča na 2*2*2=8, iz česar sledi, da se inflacijsko razmerje eksponentno povečuje s številom dimenzij. Število izpeljanih vrednosti lahko tako v primeru uporabe šestih in več dimenzij in ob visokem povprečnem razmerju na posamezno dimenzijo zelo hitro naraste čez vse razumne meje in s tem drastično zmanjša performančne karakteristike večdimenzionalne baze in nanjo vezanih aplikacij.

Rešitev problema je v uporabi manjšega števila dimenzij pri načrtovanju, saj na primer model s štirimi ali petimi dimenzijami kreira bistveno manj izpeljanih vrednosti kot model z osmimi dimenzijami. To pa vedno ni možno, saj kompleksni sistemi, ki so predmet modeliranja, pogosto zahtevajo obravnavo z večjega števila vidikov, sicer model ni popoln. Predhodna obravnava izpeljanih vrednosti je predpostavljala, da vrednosti po izračunu ostanejo shranjene v podatkovni bazi. Ob tem se pojavi vprašanje ali je sploh smiselno vse te rezultate trajno shranjevati oziroma ali ni bolje izvajati preračunavanja sproti. Seveda tudi za to ni enotnega odgovora pri čemer pa splošno vodilo pravi, da je primerno shranjevati predvsem tiste rezultate, ki se pogosto uporabljajo kot vhodni podatki drugih formul oziroma njihove formule temeljijo na velikem številu vhodnih podatkov.



Slika 6: Formula za posameznega člana dimenzije vpliva na vse celice, ki jih ta član opredeljuje

Povezave

Večdimenzionalne podatkovne baze črpajo podatke iz različnih izvorov: iz tekstovnih datotek, relacijskih podatkovnih baz, preglednic itd. V ta namen moramo tako

definirati določene povezave z zunanjim svetom, ki skrbijo za periodično obnavljanje podatkov. Z vidika procesiranja sprememb poznamo dva tipa povezav in sicer statične in dinamične. Statične povezave nimajo sposobnosti procesiranja sprememb, ki se zgodijo na izvornih podatkih, medtem ko dinamične povezave spremembe zaznajo in jih prenesejo tudi na večdimenzionalno bazo.

Drugi vidik delitve povezav se nanaša na vrsto podatkov, ki jih povezave črpajo. Skladno s tem poznamo:

- atributne povezave: črpajo podatke o lastnostih članov dimenzije
- strukturne povezave: črpajo strukturne podatke o dimenzijah, kot so podatki o članih dimenzij, hierarhijah itd.
- vsebinske povezave: črpajo vsebinske podatke, ki se shranjujejo v celice večdimenzionalne baze.

Ob navedenih obstajajo še mešani tipi povezav, ki kombinirano prenašajo različne vrste podatkov. Čeprav postopka definiranja povezav ne moremo uvrstiti med najzanimivejše dele načrtovanja in izgradnje večdimenzionalnih podatkovnih baz, pa je z vidika točnosti, ažurnosti in konsistentnosti podatkov izrednega pomena. Le dinamične, skrbno določene povezave z ustreznim izvorom (podatkovno skladišče v obliki relacijske podatkovne baze) zagotavljajo ob veliki količini podatkov pridobitev prave informacije, v pravem času, na pravem mestu.

6. Postopek načrtovanja in izgradnje večdimenzionalne podatkovne baze

Dandanes v svetu še ne obstaja neka splošno uveljavljena metodologija za izgradnjo večdimenzionalnih podatkovnih baz, kot je to v primeru informacijskih sistemov, temelječih na relacijskih podatkovnih bazah (informacijski inženiring, strukturna sistemska analiza itd). Različni avtorji v svojih delih podajajo določena priporočila, ki pa nimajo namena delovati kot metodologija, temveč bolj kot praktično vodilo za uspešno načrtovanje in izgradnjo. Navzlic temu lahko celoten proces modeliranja strnemo v naslednjih pet korakov:

- spoznavanje trenutne situacije,
- definiranje kock, dimenzij in povezav,
- definiranje hierarhij,
- definiranje članov dimenzij,
- definiranje formul.

Spoznavanje trenutne situacije

Pred samim začetkom načrtovanja večdimenzionalne podatkovne baze moramo najprej spoznati in razumeti trenutno situacijo obravnavanega sistema, za katerega želimo v končni fazi ponuditi tudi neko ustrezno rešitev. Obstoječe stanje obravnavamo z logičnega in fizičnega vidika, pri čemer v okviru fizičnega vidika proučimo predvsem obstoječe sistemske rešitve (sistem zasnovan na naboru preglednic, uporaba relacijske podatkovne baze itd.), v okviru logičnega vidika pa poslovne procese, pravila in vsebino nasploh. V pomoč so nam lahko vnaprej pripravljeni vprašalniki, seveda pa ne smemo izpustiti neposrednih intervjujev s končnimi uporabniki. Rezultat prvega koraka je logičen model sistema na visokem nivoju, kjer opredelimo vse izvore podatkov (podatkovna skladišča, preglednice, zunanji izvori itd.) in uporabnike (končni uporabniki, aplikacije). Prav tako je koristno, da že v tej fazi naredimo tudi grobo oceno prometa (število vrstic, transakcij na časovno enoto) in določimo osnovne tipe podatkov.

Definiranje kock, dimenzij in povezav

Logično načrtovanje večdimenzionalne podatkovne baze se začne z opredelitvijo njene osnovne strukture, kar vključuje obravnavo števila hiperkock (večdimenzionalna baza lahko vsebuje tudi večje število hiperkock, če orodje OLAP to omogoča) in njihovih dimenzij. V kolikor imamo že pripravljeno ustrezno podatkovno skladišče (zvezdna shema), lahko podatke o dimenzijah preprosto prenesemo v kocko z uporabo povezav. Orodja OLAP se tu razlikujejo predvsem v načinu obravnavanja spremenljivk, pri čemer nekatera definirajo posebno dimenzijo in obravnavajo spremenljivke kot člane dimenzije (TM/1), druga pa obravnavajo spremenljivke individualno (Oracle Express).

Rezultat prvega koraka logičnega načrtovanja je tako predvsem nabor dimenzij, ki opredeljujejo posamezno kocko. Pri tem si moramo prizadevati, da število dimenzij ne prekorači razumnih meja, saj ima velik vpliv na število izpeljanih vrednosti in razpršenost podatkov. Pogosto uporabljena rešitev je združevanje dimenzij, ki pa mora biti vsebinsko pravilno izvedeno, da ne pride do izgube pomembnih vidikov obravnavanega sistema.

Povezave imajo v večdimenzionalni podatkovni bazi pomembno vlogo, saj natančno opredeljujejo podatkovne tokove. Njihovo definiranje v zgodnji fazi načrtovanja ima pozitiven vpliv na vse nadaljnje modeliranje, saj omogoča testiranje pravilnega izbora dimenzij, še preden se lotimo podrobnejše specifikacije. Kasnejše spreminjanje dimenzij, ko so hierarhije, člani in spremenljivke že določene, lahko bistveno upočasni celoten projekt, saj je potrebno opraviti revizijo celotnega modela.

Definiranje hierarhij

V fazi definiranja hierarhij se osredotočimo na ugotavljanje, ali je predhodno določene dimenzije smiselno organizirati v hierarhije. V primeru pozitivnega odgovora nadalje analiziramo strukturo hierarhij (simetrična ali asimetrična), določimo število članov na posamezni

ravni, ravni poimenujemo itd. Posamezna dimenzija ima lahko tudi več hierarhij s skupnim ali različnimi korenskimi elementi, kar omogoča združevanje članov po različnih kriterijih in zagotavlja povsem nove poglede na obravnavano tematiko. Pri definiranju hierarhij moramo posebno pozornost posvetiti analizi, ali je vsak temeljni član dimenzije skozi hierarhično drevo povezan s korenom. Pogosto se namreč dogaja, da se posamezne povezave sčasoma zaradi različnih reorganizacij porazgubijo, posledica pa je nepopolnost in netočnost izkazanih rezultatov. Z definiranjem hierarhij zaključimo logično načrtovanje večdimenzionalne podatkovne baze na višjem nivoju.

Definiranje članov dimenzij

Definiranje članov dimenzij je prvi korak v izgradnji podrobnega logičnega modela. Le-ta je sestavljen iz nabora dekompozicijskih diagramov, ki natančno specificirajo razporeditev članov dimenzije po hierarhičnih ravneh. Problem nastane, če dimenzije vsebujejo veliko število članov, kar se v realnem svetu pogosto dogaja. V tem primeru je nemogoče, pa tudi nesmiselno vse člane prikazati na diagramu, zato se omejimo zgolj na dovolj reprezentativen vzorec, ki zagotavlja pravilno razumevanje dimenzijske strukture.

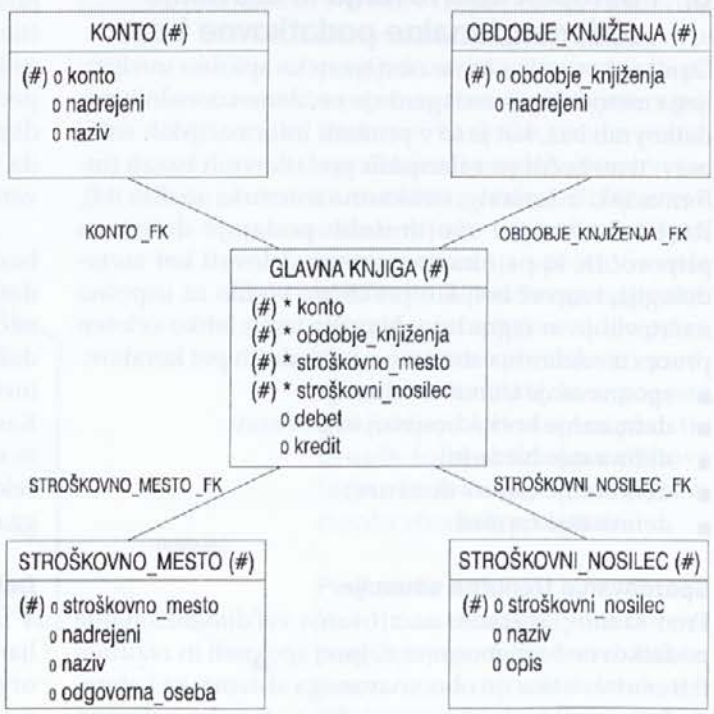
Definiranje formul

Rezultat dosedanjega postopka je dovolj podrobno specificirana podatkovna struktura večdimenzionalne baze. Obstaja pa še vedno odprta naloga definiranja formul, pri čemer se moramo takoj na začetku vprašati, kakšen je namen naše programske rešitve in v povezavi s tem, kakšne tipe formul bomo uporabili. Formule glede na njihovo namembnost namreč delimo na opisne, razlagalne, napovedovalne in povzemajoče. Najbolj kompleksne, napovedovalne, se uporabljajo za ugotavljanje različnih možnih situacij v prihodnosti in so zato dobra podlaga za sprejemanje odločitev o bodočem poslovanju.

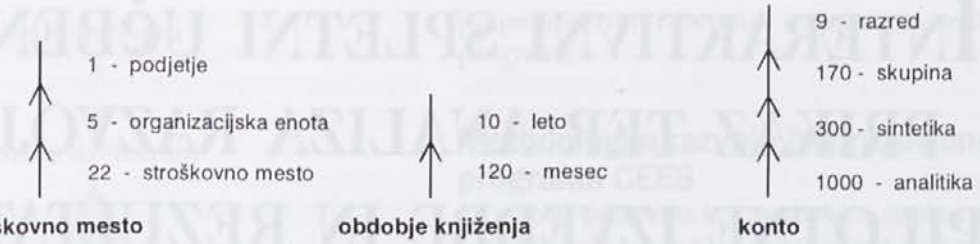
Napovedovalne formule temeljijo na povzemajočih, katerih osnovna naloga je izračunavanje vrednosti za področja, kjer meritve še niso bile opravljene. Večinoma se uporabljajo za napovedovanje prihodnosti, najdemo pa jih tudi pri obravnavi sedanjosti in preteklosti, če določenih meritev ni oziroma ni bilo možno izvesti. Povzemajoče formule so kombinacija znanih odvisnosti med podatki in verjetnostnih funkcij njihovega spreminjanja skozi čas, temeljijo pa na razlagalnih formulah. Razlagalne formule predstavljajo relacije med obstoječimi podatki, kot na primer razmerje med dejansko in načrtovano prodajo, saldo (razlika med debetom in kreditom) itd. Najpreprostejše formule, imenovane opisne formule, pa se večinoma uporabljajo posredno v okviru definiranja hierarhij za agregacije in povprečja. Vsaka zahtevnejša večdimenzionalna podatkovna baza vsebuje vsaj opisne in razlagalne formule, pri čemer pa pravo vrednost pridobi šele z uporabo povzemajočih in napovedovalnih.

7. Primer glavne knjige

Za zaključek si oglejmo še zgled načrtovanja večdimenzionalne podatkovne baze za potrebe stroškovnega računovodstva oziroma glavne knjige. V transakcijskem sistemu je glavna knjiga ponavadi predstavljena z eno samo, daljšo tabelo, katere posamezni zapisi (knjiženja) so enolično določeni s številko temeljnice, s katero so knjiženi, postavko temeljnice in letom. Drugi pomembnejši atributi so še obdobje knjiženja, konto, stroškovno mesto, stroškovni nosilec, vrsta stroška, proti konto, datum knjiženja, vrsta dokumenta, številka dokumenta in seveda debet ter kredit. V fazi izgradnje področnega skladišča podatkov ali samostojnega data marta preoblikujemo tabelo glavne knjige na način, ki kar najbolj ustreza zahtevam programskih orodij OLAP in namenom naše programske rešitve, kar pomeni, da moramo že tedaj izvesti tudi analizo obstoječe situacije in ugotoviti prednostne cilje. Pri izgradnji podatkovnega skladišča se poslužujemo različnih tehnik kot so izločevanje ali združevanje atributov, agregiranje podatkov itd. Slika 8 prikazuje področno skladišče podatkov v obliki zvezdne sheme za glavno knjigo, pri čemer



Slika 8.: Zvezdna shema podatkovnega skladišča za glavno knjigo



Slika 9: Večdimenzionalna podatkovna struktura za glavno knjigo

so že upoštevane značilnosti končne rešitve, ki bo obravnavala podatke s štirih različnih vidikov: z vidika obdobja knjiženja, stroškovnega mesta, stroškovnega nosilca in konta.

Ustrezno definirano skladišče podatkov zagotavlja dobro izhodišče za nadaljnje načrtovanje večdimenzionalne podatkovne baze. Že sama zvezdna shema (ena tabela dejstev in štiri dimenzijske tabele) nakazuje, da bo večdimenzionalna baza sestavljena iz ene hiperkocke s štirimi dimenzijami: obdobje knjiženja, stroškovno mesto, stroškovni nosilec in konto. Nadalje je iz sheme razvidno, da lahko definiramo atributne in strukturne povezave z vsemi štirimi dimenzijskimi tabelami, vsebinske pa s tabelo dejstev.

Definiranje hierarhij za posamezno dimenzijo je pogojeno z obstoječo organizacijo podatkov. V našem primeru imamo tri nivoje stroškovnih mest (podjetje, organizacijska enota in stroškovno mesto), štirinivojski kontni plan (razred, skupina, sintetika in analitika), in dvonivojsko časovno dimenzijo obdobje knjiženja (leto, mesec). Ob opredelitvi hierarhij izdelamo še oceno števila članov na posameznih nivojih, kar nam omogoča izračun inflacijskega razmerja in kompleksnosti večdimenzionalne podatkovne baze. Preostane nam le še določitev članov dimenzij in formul. Člane dimenzij preberemo iz obstoječih dimenzijskih tabel, opisne formule se generirajo posredno pri definiranju hierarhij, kot razlagalno formulo pa lahko določimo saldo (debet-kredit). Na ta način smo zaključili izgradnjo osnovne podatkovne strukture, ki predstavlja primeren temelj za nadaljnje definiranje kompleksnejših povzemajočih in razlagalnih formul.

8. Zaključek

Direktorski informacijski sistemi, zasnovani na tehnologiji večdimenzionalnih podatkovnih baz in podatkovnih skladišč, so že danes v svetu sestavni del informacijskih sistemov vseh pomembnejših podjetij. V Sloveniji se dandanes še vedno večinoma ukvarjamo z informatizacijo poslovnih procesov (transakcijskih sistemov), ob tem pa je izgradnja podatkovnih skladišč zapostavljena. Se pa vse več vodilnih zaveda pomena, ki jih ima prava informacija v poplavi velike količine raznovrstnih podatkov, kar potrjujejo projekti, ki se trenutno izvajajo v nekaterih večjih slovenskih podjetjih in bankah. Orodja OLAP omogočajo izgradnjo učinkovitih programskih rešitev v ta namen, pri razvoju katerih ima pomembno vlogo načrtovanje večdimenzionalnih podatkovnih baz. Pričakovati je, da bo ta segment informacijske tehnologije v prihodnosti še pridobil na pomenu ter postal standardna nadgradnja relacijskih in objektnih podatkovnih baz.

Literatura

1. Inmon William: Building The Data Warehouse, John Willey²&Sons, 1996.
2. Ralph Konball: The Data Warehouse Toolkit, John Willey²&Sons, 1996.
3. E. Thomsen: OLAP Solutions - Building Multidimensional Information Systems, Wiley Computer Publishing, 1997.
4. M. J. Corey, M. Abbey, I. Abramson, B. Taub: Oracle8 Data Warehousing, Oracle PressÔ, Osborne/McGraw-Hill, 1998.
5. Len Silverston, W. H. Inmon, Kent Graziano: The Fata Model Resource Book, Wiley Computer Publishing, 1997.

◆

Mag. Sebastian Lahajnar je diplomiral leta 1997 na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Po diplomi je vpisal podiplomski študij na Ekonomski fakulteti, smer Informacijsko upravljalske vede, in leta 1999 zagovarjal magistrsko delo. Zaposlen je kot razvijalec v podjetju PRIS Inženiring, kjer se ukvarja z razvojem poslovnih informacijskih sistemov.

◆

Alenka Rožanec je leta 1991 končala Srednjo šolo za računalništvo in leta 1997 diplomirala na Fakulteti za računalništvo in informatiko Univerze v Ljubljani. Leta 1998 je vpisala podiplomski študij Informacijskih sistemov in odločanja na omenjeni fakulteti. Zaposlena je kot informatik na Upravni enoti Ljubljana.

◆