# Analiza in primerjava pristopov pri gradnji podatkovnih skladišč

Izidor Golob, Tatjana Welzer, Boštjan Brumen
Fakulteta za elektrotehniko, računalništvo in informatiko
Univerza v Mariboru, Smetanova 17, 2000 Maribor
{izidor.golob, welzer, brumen}@uni-mb.si

## Izvleček

Podatkovno skladiščenje s svojimi strukturami in procesi, prirejenimi v podporo poslovnemu procesu, ter podatki, prečiščenimi in integriranimi v procesu migracije podatkov, omogoča celovit pogled na podatke posamezne organizacije. V prispevku analiziramo in primerjamo tri temeljne pristope pri gradnji podatkovnih skladišč. Glede na izbrani pristop je rezultat centralizirana, porazdeljena ali federativna zgradba podatkovnega skladišča. Rezultati primerjave kažejo, da se pristopi bistveno razlikujejo po osnovnih sestavnih elementih, uporabljenem podatkovnem modelu in metodologiji izgradnje. Na podlagi rezultatov podajamo smernice za izbiro optimalnega pristopa pri gradnji podatkovnega skladišča za določeno organizacijo. Splošna najboljša rešitev ne obstaja, saj ima vsak pristop svoje posebnosti, tako dobre kot slabe.

## Abstract

**Analysis and Comparison of Approaches to Building Data Warehouses**

Data warehousing along with its structures and processes, is designed to support the business process. A data warehouse, containing data, cleansed and integrated in the data migration process, gives an integrated view on the enterprise's business data. In this paper three basic approaches to building data warehouse are analyzed and compared. Depending on the approach chosen, the result is centralized, distributed or federated data warehouse architecture. The results show that there is a fundamental difference among the presented approaches due to differences in concepts, data models and methodology used. Based on the results, a method for determining an optimal approach to building data warehouse for given enterprise is given. In general, the perfect approach does not exist as each one has its own strengths and weaknesses.

## 1 Uvod

**Pri upravljanju informacij podjetja ločimo dve vrsti računalniških sistemov: za sprotno obdelavo transakcij (OLTP, angl. on-line transaction processing) in analiziranje. Sistemi za sprotno obdelavo transakcij, ali krajše transakcijski sistemi, podpirajo procese z značilnimi transakcijskimi procesnimi sistemi, ki s svojimi operativnimi podatkovnimi bazami zajemajo transakcije – poslovne dogodke. Sistemi za analiziranje so povezani z analizo informacij o osnovnih procesih in njihovim nadzorom s sredstvi sistemov za upravljanje informacij.**

Osnovna značilnost operativnih podatkovnih baz iz sistemov za sprotno obdelavo transakcij so podrobni atomarni podatki, shranjeni v stabilnih normaliziranih podatkovnih strukturah, optimiziranih za zajem transakcij ter vpis in ažuriranje podatkov [Barquín 1997; Anahory 1997; Bischoff, 1997]. Operativni sistemi in sistemi podatkovnih skladišč nosijo številne nasprotne značilnosti. Če jim pridružimo še omejitve strojne in programske opreme, operativne podatkovne baze niso dovolj prilagodljive hitro se spreminjajočim zahtevam uporabnikov (največkrat so to vodilni delavci – poslovodstvo, upravljavci ali analitiki), ki želijo informacije za sprejemanje odločitev v sprejemljivem času. Prav podatkovno skladiščenje (angl. data warehousing) s svojimi strukturami in procesi, prirejenimi v podporo poslovnemu procesu (podatkovna skladišča so oblikovana za kompleksna ad hoc povpraševanja) ter podatki, prečiščenimi in integriranimi v procesu migracije podatkov, omogoča tistim, ki odločajo in skrbijo za razvoj podjetja, celovit pogled na podatke posamezne organizacije ne glede na uporabljene strojne in programske rešitve v posameznih operativnih okoljih. V sinergiji z dodatnimi programskimi analitičnimi orodji, izmed katerih je še posebej pomembna sprotna analitična obdelava (angl. On-Line Analytical Processing –

OLAP), ter podatkovnim rudarjenjem predstavljajo vrh današnje informacijske podpore, saj omogočajo celovit pogled na poslovanje organizacije skozi različne vidike.

Pomen podatkovnega skladiščenja, dosedanje uspehe in pripravljenost investitorjev v še večje investicije potrjujejo tudi zadnja poročila, ki napovedujejo rast investicij iz 37,4 milijard USD, kolikor so znašale investicije v podatkovno skladiščenje v svetovnem merilu leta 1999, na 150 milijard USD leta 2003 [Hammond 2000], kar predstavlja 40-odstotno letno rast.

Podatkovno skladiščenje kljub uspešnim izdelkom še ni zrela disciplina. Predvsem zaradi napredkov v strojni in programski opremi se nenehno pojavljajo nove rešitve in zgradbe. Ena izmed najpomembnejših odločitev, s katero se mora soočiti vsak načrtovalec podatkovnega skladišča že na začetku, je izbira primernega pristopa pri gradnji podatkovnega skladišča. Izbira pristopa je kritična, saj le-ta določa podatkovni model, vlogo področnih skladišč ter sosledje korakov v razvojnem ciklu. Glede na izbran pristop je rezultat centralizirana, porazdeljena ali federativna zgradba podatkovnega skladišča.

Kljub navedenemu je prav na področju poznavanja in razumevanja pristopov pri gradnji podatkovnih skladišč največ zmede; dodatno jo povzročajo še avtorji in zagovorniki posameznih pristopov s svojimi opozorili o neprimernosti ostalih. Cena projekta podatkovnega skladiščenja je v primerjavi z ostalimi projekti na področju informatike visoka, saj vključuje poznavanje podrobnosti obstoječih in novih sistemov. Zato je strah pred nepravilno izbiro pristopa utemeljen.

### 1.1 Metodologija raziskave

Preliminarna raziskava je pokazala, da zaradi razlik med poimenovanji, med razvojnimi cikli in med izhodišči neposredna primerjava pristopov pri gradnji podatkovnih skladišč, ki rezultirajo v različnih zgradbah podatkovnih skladišč, ni možna. Zato je po uvodni predstavitvi izvedena identifikacija stičnih točk pristopov. Le na podlagi identificiranih medsebojno primerljivih elementov je namreč možno izvesti veljavno primerjalno analizo. Na podlagi njenih rezultatov so identificirane želene in neželene lastnosti posameznih pristopov in iz njih izhajajočih zgradb. Ugotovitve so strnjene v tabelo, ki predstavlja metodo za določitev optimalnega pristopa pri gradnji podatkovnega skladišča za posamezno organizacijo glede na parametre.

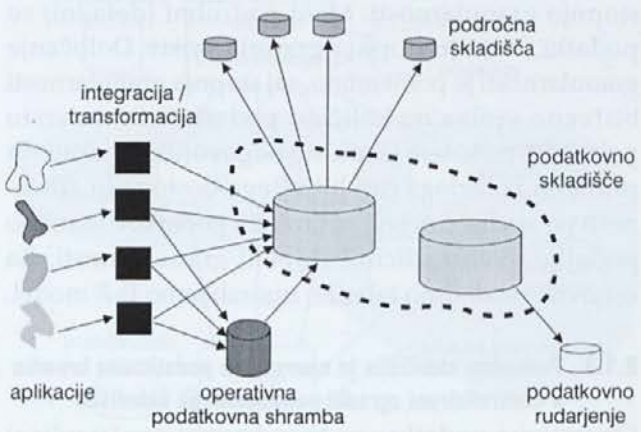
## 2 Pristopi pri gradnji podatkovnih skladišč

### 2.1 Centralizirani pristop

V središču centralizirane zgradbe, ki je rezultat uporabe centraliziranega pristopa gradnje podatkovnega skladišča je podatkovno skladišče zaključenega organiziranega sistema. Le-to »hrani« področna skladišča, polni pa se iz operativnih podatkovnih baz ter operativnega podatkovnega skladišča (slika 1). Največji zagovornik takšne zgradbe je Inmon [Inmon 1996, 1999b]. Strogo rečeno so v taki zgradbi področna skladišča odvisna struktura, saj so podatki pridobljeni oz. naloženi izključno iz podatkovnega skladišča organizacije.

#### 2.1.1 Načrtovanje in gradnja centraliziranega podatkovnega skladišča

Razlike med operativnim svetom in svetom podatkovnega skladišča so izrazito vidne tudi iz opisa razvojnih ciklov. Operativno okolje je podprto s klasičnim razvojnim ciklom, ki je vodeno s strani zahtev. Tako je potrebno najprej razumeti zahteve, šele nato preidemo v faze načrtovanja in razvoja. Razvojni cikel podatkovnega skladišča je podatkovno voden, saj pričnemo s podatki. Po integraciji podatkov pogledamo, če je potrebno njihovo dodatno uglaševanje in to po potrebi tudi storimo. Rezultati programov so analizirani in šele na koncu razumemo




Slika 1: **Centralizirano podatkovno skladišče**

zahteve. Vrstni red posameznih faz razvojnega cikla operativnega okolja in okolja podatkovnega skladišča je popolnoma obrnjen.

### 2.1.2 Podatkovni model v centralizirani zgradbi podatkovnega skladišča

Struktura podatkovnega skladišča je normalizirana. V maloštevilnih primerih je struktura le delno denormalizirana. Delno denormaliziranost, torej delno odstopanje od zahtev po doseganju tretje normalne oblike, lahko uvedemo v naslednjih primerih:

- Kjer je znano, da se bodo redundantni podatki redno uporabljali skupaj z drugimi podatki in zato dopuščamo redundantnost.
- Kjer so enkrat izračunani podatki večkrat uporabljeni (npr. shranimo letno bruto plačo, čeprav je to izpeljan, izračunljiv podatek).
- Če sklepamo, da bodo skupine podatkov normalno in pogosto uporabljene skupaj, formiramo zanje nov skupen prostor. Tako lahko npr. mesečne podatke za mesece januar, februar itn. fizično nakopičimo na eno samo lokacijo in s tem poenostavimo in povečamo hitrost fizičnega dostopa.
- Če sklepamo, da se verjetnost dostopa do posameznih podatkovnih elementov (atributov) pri uporabi bistveno razlikuje in zato izvedemo ločitev podatkov.

Kljub morebitni denormaliziranosti strukture oziroma podatkovnega modela podatki obdržijo značilnost močne normaliziranosti, saj delna denormalizacija ne odraža zahtev posameznega oddelka, temveč le izboljša učinkovitost vsem uporabnikom.

Podatki v podatkovnem skladišču imajo različno stopnjo granularnosti. Manj podrobni (detajlni) so podatki, višja je stopnja granularnosti. Določanje granularnosti je pomembno, saj stopnja granularnosti bistveno vpliva na količino podatkov in na vrsto poizvedb, na katere je možno odgovoriti. V nekaterih primerih iz razloga čim hitrejšega dostopa in zmožnosti po analizi čim bolj podrobnih podatkov hranimo podatke dveh različnih stopenj granularnosti. Za osnovno modelirno tehniko uporabljamo E-R model.

### 2.1.3 Področno skladišče in operativna podatkovna hramba v centralizirani zgradbi podatkovnega skladišča

Operativna podatkovna hramba je v predstavljeni strukturi eden izmed podatkovnih virov podatkovnega skladišča. Je hibridna struktura, ki izpolnjuje tako operativne kot tudi analitične zahteve: zagotavlja majhen transakcijski odzivni čas, hkrati pa je tudi prostor integriranih podatkov. V primerjavi s podatkovnim skladiščem izpolnjuje zahteve po predmetni usmerjenosti in integriranosti, ne vsebuje pa zgodovinskih podatkov, temveč le trenutne, aktualne in detajlne podatke določene organizacije. Nasprotno kot podatkovno skladišče pa je operativna podatkovna hramba lahko ažurirana na podlagi transakcij [Inmon 1999b].

Izhajajoč iz zgradbe centraliziranega podatkovnega skladišča (gl. sliko 1) je osrednje podatkovno skladišče edini vir podatkov za področno skladišče. Področno skladišče ne ločuje med podatki, ki so prišli v podatkovno skladišče neposredno iz operativnega sveta ali preko operativne podatkovne hrambe.

Osnovne značilnosti, ki ločijo področno skladišče in podatkovno skladišče so naslednje:

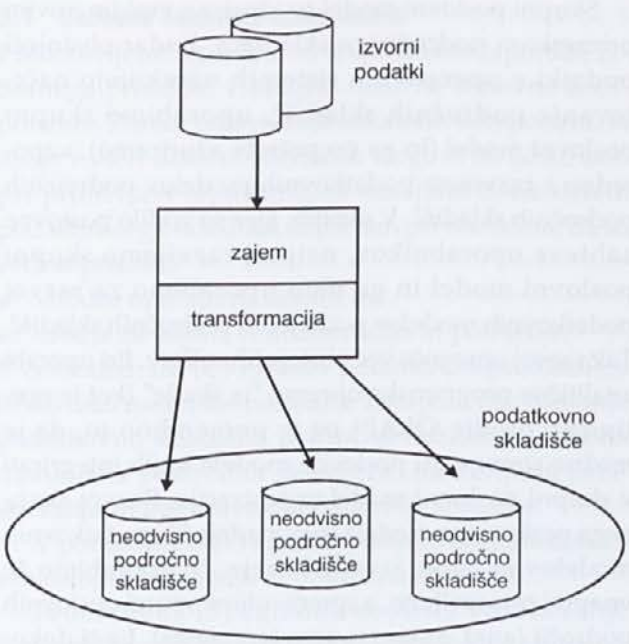
- podatkovno skladišče vsebuje veliko količino zelo podrobnih podatkov iz daljšega obdobja (npr. 10 let) v enostavnih strukturah, področno skladišče pa le podatke omejene vsebine, s takšno stopnjo granularnosti in iz takega časovnega obdobja, kot narekujejo potrebe oddelka [Inmon 1999a];
- strukture podatkovnega skladišča so namenjene neznani uporabi, strukture področnega skladišča pa so načrtovane za specifične, znane namene;
- področna skladišča so manjša in
- podatkovno skladišče vsebuje tako podatke nižje granularnosti kot tudi sumarne podatke poslovanja celotne organizacije.

## 2.2 Pristop gradnje porazdeljene zgradbe podatkovnega skladišča

Področno skladišče je podmnožica podatkovnega skladišča določene organizacije. V porazdeljeni zgradbi je podatkovno skladišče le unija področnih skladišč. Področno skladišče igra ponavadi vlogo oddelčnega, krajevnega ali funkcionalnega podatkovnega skladišča in podpira eno ali več specifičnih področij.

Tipično porazdeljeno zgradbo, katere največji zagovornik je Kimball [Kimball 1998], prikazuje slika 2.

Organizacija kot del iterativnega procesa gradnje podatkovnega skladišča zgradi vrsto porazdeljenih področnih skladišč in jih na koncu poveže v logično podatkovno skladišče celotne organizacije. Hackney tak pristop brez zadržkov poimenuje "od zgoraj navzdol" [Hackney 2000a].




Slika 2: **Porazdeljeno podatkovno skladišče**

Področna skladišča postavljajo specifične oblikovalske zahteve. Vsako področno skladišče mora biti predstavljeno z dimenzijskim modelom, ki mora biti znotraj enotnega podatkovnega skladišča skladen.

Skladna dimenzija (angl. conformed dimension) je dimenzija, za katero je značilno, da ima enoličen pomen, ne glede na to, s katero tabelo dejstev jo povežemo. Zagotavlja tudi, da je podatek predstavljen le enkrat. Glavna naloga skupine načrtovalcev podatkovnega skladišča pri načrtovanju porazdeljene zgradbe podatkovnega skladišča je vzpostavitev, objava in vzdrževanje skladnih dimenzij, kot tudi zagotavljanje njihove dosledne uporabe. Brez upoštevanja koncepta skladnih dimenzij podatkovno skladišče ne more delovati kot integrirana celota.

### 2.2.1 Podatkovni model v porazdeljeni zgradbi podatkovnega skladišča

Struktura področnih skladišč je denormalizirana, v določenih primerih le delno normalizirana. Osnovni podatkovni model je dimenzijski, za osnovno modelirno tehniko pa uporabljamo dimenzijsko modeliranje.

Dimenzije, še posebej skladne, imajo navadno atomarne (granularne) podatke. To pomeni, da morajo biti tudi osnovne tabele dejstev na najnižjem nivoju, ki obstaja med pripadajočimi dimenzijami. To dejstvo olajšuje prehod podatkov iz operativnih podatkovnih baz v tabele dejstev.

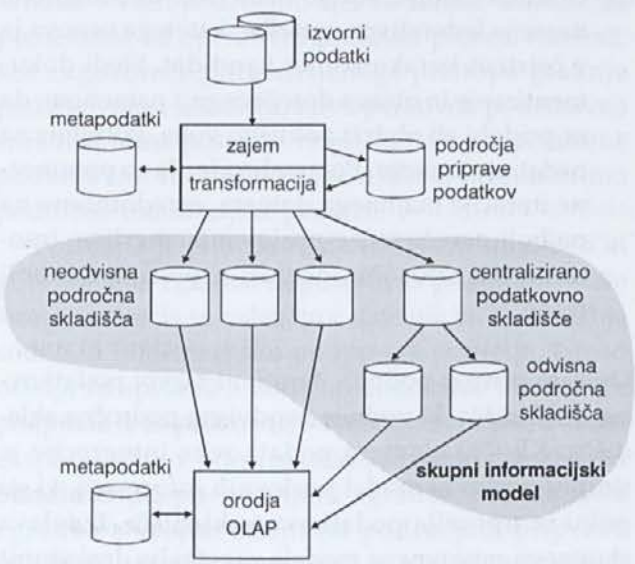
### 2.2.2. Ostale posebnosti

Zgradba omogoča razmeroma hitro gradnjo prvega področnega skladišča, ki ima visok poslovni (pozitivni) vpliv oziroma pomen, hkrati pa jo je razmeroma enostavno implementirati. Potrebe organizacije po analizah so razvidne iz poslovnih zahtev, od koder nato izhaja tudi določitev prioritet. Tak pristop je zelo pomemben, saj v najkrajšem času pridobimo delujoče podatkovno skladišče, s tem pa podporo zagovornikov s strani vodstva in uporabnikov. Le-to omogoča naslednjo iteracijo, t. j. gradnjo novega področnega skladišča.

## 2.3 Federativni pristop h gradnji podatkovnega skladišča

Federativno podatkovno skladišče (angl. federated data warehouse) je hibridna rešitev, temelječa na skupnem poslovnem modelu (angl. common business model) in področjih priprave informacij (angl. information staging areas), ki so v skupni rabi [White 2000; Hackney 2000a, b, c, d] (slika 3).

Takšna zgradba zagotavlja nizke stroške in hitro povrnitev vloženih sredstev z uporabo neodvisnih področnih skladišč, pri čemer kasnejša podatkovna integracija ni potrebna.




Slika 3: **Federativno podatkovno skladišče**

### 2.3.1 Gradnja federativnega podatkovnega skladišča

Gradnjo federativnega podatkovnega skladišča bi lahko strnili v šest korakov [Hackney 2000b]:

(1) Dokumentiranje obstoječih sistemov podatkovnih in področnih skladišč rezultira v entitetnem diagramu, ki prikazuje sisteme in vse podatkovne tokove med njimi, vključno s tokom meta podatkov.

(2) Dokumentiranje obstoječih sistemov na nivoju toka podatkov vključuje podatkovni tok, pripadajoče korake transformacije in integracije ter repozitorije meta podatkov. Vsak podatkovni element mora biti ocenjen v smislu kakovosti, razpoložljivosti in enostavnosti dostopa.

(3) Določitev podatkov, ki prinašajo dodano vrednost in imajo dovolj visok pomen oziroma vpliv v celotnem sistemu. V tem koraku se išče posebna, najbolj pomembna podatkovna integracija.

(4) Zbiranje kandidatov iz prejšnjega koraka in analiziranje njihovega vpliva in možnosti za implementacijo. V tem koraku so tudi izbrani ustrezni kandidati z najnižjo stopnjo tveganja, taki, ki tudi največ prispevajo k izpolnitvi strateškega načrta organizacije. Pomembno je, da so izpuščeni tisti, ki so zanimivi le za ožji krog (čeprav poslovnih) uporabnikov.

(5) Implementacija orodja za zajem, transformacijo in polnjenje podatkov, ki podpira skupen, globalni repozitorij metapodatkov sistemov podatkovnih in področnih skladišč.

(6) Gradnja manjše, strogo namenske in usmerjene iteracije federativne zgradbe, katerega osnova je v četrtem koraku izbran kandidat. Sledi dokumentiranje in objava doseženega z namenom, da se pridobi ali obdrži politično volja, potrebna za nadaljnje iteracije. Pomembno je, da so posamezne iteracije majhnega dometa, osredotočene na najbolj pereče točke poslovanja, merljive (moramo biti sposobni oceniti uspeh) in čimbolj tržne.

Opisani pristop podpira iterativni razvoj podatkovnega skladišča, ki vsebuje neodvisna področna skladišča. Ključni element podatkovne integracije je skupni poslovni model poslovnih informacij, ki ga polni in upravlja podatkovno skladišče. Izdelava skupnega poslovnega modela zagotavlja doslednost pri uporabi imen podatkov in poslovnih definicij v vseh procesih podatkovnega skladišča.

Skupni poslovni model se ažurira z vsakim novim primerkom področnega skladišča. Kadar obstoječi podatki v operativnih sistemih narekujejo načrtovanje področnih skladišč, uporabimo skupni poslovni model (in ga po potrebi ažuriramo), vzporedno z razvojem podatkovnih modelov podrejenih področnih skladišč. V okoljih, kjer so vodilo poslovne zahteve uporabnikov, najprej razvijemo skupni poslovni model in ga nato uporabimo za razvoj podatkovnih modelov podrejenih področnih skladišč. Tak razvoj omogoča več obstoječih rešitev. Pri uporabi analitične programske opreme "iz škatle" (kot je npr. tipično orodje OLAP) pa je pomembno to, da je možno spremeniti poslovne modele in jih integrirati v skupni poslovni model organizacije. Razvoj skupnega poslovnega modela in pripadajočih podatkovnih modelov se lahko izvede hitreje, če uporabimo že vnaprej pripravljene, a spremenljive vzorci poslovnih področij (angl. business area templates). Le-ti dokumentirajo poslovne metrike in pravila, ki se uporabljajo pri analizi in modeliranju poslovnega procesa.

Ob dodajanju novih področnih skladišč v sistem podatkovnega skladišča se obstoječe rutine za zajem in podatki v področjih priprave po potrebi ponovno uporabijo oziroma izboljšajo. Tak način še posebej dobro deluje v federativnem podatkovnem skladišču, kjer lahko skupni poslovni model uporabimo pri načrtovanju področij priprave in programskih rutin za zajem podatkov.

## 3 Primerjava pristopov

Za določitev učinkovitega in ustreznega pristopa pri gradnji podatkovnega skladišča, katerega rezultat je zagotovljen najmanjši odzivni čas in učinkovita izraba virov, je nujno potrebno poznavanje lastnosti posameznih pristopov, pozitivnih in negativnih. Pravilna odločitev zmanjša tveganje pri novih projektih podatkovnega skladiščenja (po nekaterih virih [Wells 2000] propade kar 70 % projektov podatkovnih skladišč), prispeva k izboljšanju obstoječih podatkovnih skladišč z novimi področnimi skladišči zaradi odprtosti sistemov ter omogoča prilagoditev preveč centraliziranih podatkovnih skladišč, ki ne morejo učinkovito delovati v nehierarhičnem okolju, na bolj federativno zgradbo.

Pri primerjavi pristopov smo zajeli tiste elemente, ki določajo zgradbo: osnovne komponente, uporabljen podatkovni model, razvojni cikel ter zgradbo meta podatkov in način njihovega upravljanja.

### 3.1 Osnovne sestavne komponente

Predstavljene zgradbe, ki so posledica uporabe izbranega pristopa, vsebujejo različne osnovne komponente. Zaradi prekrivanja nekaterih komponent, ki imajo v vseh sistemih identično vlogo in jih lahko zato pri primerjavi izpustimo, se omejimo le na sistem podatkovnega skladišča in pri tem privzamemo, da so vedno prisotni:

- vhodni operativni sistemi ter
- orodja za zajem, transformacijo in polnjenje.

V centralizirani (C) zgradbi podatkovnega skladišča lahko tako nastopajo naslednje komponente: osrednje podatkovno skladišče, področna skladišča, posebna namenska področna skladišča, namenjena raziskovanju in operativne podatkovne shrambe.

V porazdeljeni (P) zgradbi podatkovnega skladišča samostojno nastopajo le področne shrambe.

Federativna (F) zgradba dopušča združevanje samostojnega podatkovnega skladišča ali sistemov podatkovnih skladišč, področnih shramb, orodij za podatkovno rudarjenje in analitičnih aplikacij, katerih tipični predstavniki so orodja OLAP.

Tabela 1 prikazuje možne udeležene samostojne komponente v posameznih zgradbah:

#### 3.1.1 Diskusija

Pomembna razlika med pristopi je v pojmovanju vloge področnega skladišča. V centralizirani zgradbi je področno skladišče odvisno (izpeljano) iz osrednjega, podatkovnega skladišča. To je hkrati največja moč področnega skladišča in hkrati njegova najšibkejša točka. Moč zato, ker jih zgradimo dokaj enostavno, saj podatke le izpeljemo iz osrednjega skladišča, kar pogosto rezultira v njihovem (pre)velikem številu. Prav vzdrževanje večjega števila strukturnih elementov je šibka točka centralizirane zgradbe. Prav tako zaradi svoje odvisnosti od podatkovnega skladišča zaključenega organiziranega sistema (ZOS), področnega skladišča ne moremo izgraditi, preden ni podatkovno skladišče ne samo načrtovano in izdelano, ampak tudi implementirano.

| Pristop / Konstrukt | C | P | F |
|---|---|---|---|
| Podatkovno skladišče | DA, osrednje | NE | NE |
| Področno skladišče | DA | DA | DA |
| Operativna podatkovna shramba | DA | NE (je vgrajena v sistem porazdeljene zgradbe) | DA |
| Ostali | NE | NE | DA |

Tabela 1: **Primerjava udeleženih samostojnih komponent**

Zaradi zmanjšane količine podatkov, ki jo je potrebno predelati pri povpraševanjih, so lahko področna skladišča ugodnejša rešitev kot podatkovna. Vendar je težko zagotoviti, da nobeno izmed povpraševanj nad izbranim področnim skladiščem ne bo zahtevalo podatka, ki ga v področnem skladišču ni. Področna skladišča so po definiciji manjša, v centralizirani zgradbi pa tudi praviloma bolj denormalizirana. V primerjavi s podatkovnimi skladišči jih zato hitreje zgradimo, jih lažje upravljamo, pa tudi cena upravljanja je nižja. Kljub vsemu navedenemu zasledimo v virih mnogo opozoril, da so samostojna oziroma neodvisna področna skladišča le "metanje peska v oči" in so "izum podjetij, ki bi rada na hitro zaslužila". Izkušnje so pokazale, da samostojna področna skladišča niso rešitev problemov in vodijo v daljšem časovnem obdobju v še večjo množico samostojnih, neintegriranih podatkovnih baz.

Med posameznimi pristopi prihaja znova do velikih razlik pri vlogi in razumevanju operativne podatkovne shrambe. Inmonova definicija pravi, da je to integrirana, nestanovitna, vendar do minute verna slika poslovnega procesa. Takšna struktura je med drugim uporabna pri trženju in stikih s strankami, skratka v vseh področjih, kjer so zadnje transakcije pomembne za operativni poslovni proces. Kimball, kot zagovornik porazdeljenega pristopa gradnje podatkovnega skladišča, kjer operativna podatkovna shramba nima posebne vloge, označuje operativne podatkovne shrambe kot strukture, kjer hranimo podrobne transakcijske podatke.

Potrebno je poudariti, da tudi centralizirana in porazdeljena zgradba dopuščata in podpirata ostale komponente, ki se nahajajo v sistemih za analiziranje sodobno informacijsko podprtega podjetja, kot so orodja za sprotno analitično obdelavo. Vendar jih podpirata le implicitno ali v najboljšem primeru delno eksplicitno kot v primeru orodij za sprotno analitično obdelavo pri porazdeljeni zgradbi. Federativna zgradba s svojim skupnim področjem priprave podatkov eksplicitno podpira in zato vključuje tudi ostale komponente.

### 3.2 Podatkovni model

Izjemno pomembna je uporaba različnih podatkovnih modelov, saj izbira podatkovnega modela bistveno vpliva na načrtovanje podatkovne baze sistema podatkovnega skladišča. Podatkovni model določa tudi vsebino in strukturo podatkovne baze podatkovnega ali področnega skladišča. Najpomembneje je, da zagotavlja uporabnikom sprejemljivi odzivni čas.

Tabela 2 prikazuje možne podatkovne modele pri posameznih pristopih:

| Model \ Pristop | C | P | F |
|---|---|---|---|
| Normalizirani (E-R) | DA | NE | DA |
| Denormalizirani (dimenzijski) | DA, vendar le za področna skladišča | DA | DA |
| Drugi (npr. objektni) | NE | NE | DA |

Tabela 2: **Podatkovni model v posameznih pristopih (primerjalno)**

#### 3.2.1. Diskusija

Zagovorniki centralizirane zgradbe, po njenem največjem zagovorniku imenovani tudi »inmonisti«, trdijo, da mora biti podatkovno skladišče razvito z uporabo E-R modela, ker so normalizirani podatki idealna struktura podatkovnega skladišča. Vendar je za gradnjo področnega skladišča dovoljena in celo priporočena tudi uporaba dimenzijskega modeliranja. Nasprotno pa zagovorniki porazdeljene zgradbe na čelu s Kimballom verjamejo, da je podatkovno skladišče možno modelirati izključno z dimenzijskim modeliranjem oziroma zvezdasto shemo.

Uporaba dimenzijskega modeliranja - zvezdaste sheme pomeni boljše razumevanje in boljšo učinkovitost glede na E-R model, njegova uporaba pa naj ne bi prinašala nobene izgube informacij. Vsak E-R model podatkovnega skladišča je lahko namreč predstavljen kot množica zvezdastih shem, in to brez izgube informacij [Kimball 1996; Firestone 1998[.

Federativna zgradba dopušča obe, normalizirano in denormalizirano strukturo.

### 3.3 Razvojni cikel

Izbira pristopa pri gradnji podatkovnega skladišča vpliva tudi na razvojni cikel. V tabeli 3 so primerjalno prikazani razvojni cikli posameznih pristopov.

#### 3.3.1 Diskusija

Ugotovimo lahko, da se predlagani razvojni cikli vseh predstavljenih pristopov bistveno razlikujejo med seboj in se razlikujejo tudi od tradicionalnega razvojnega cikla.

Sistemi, razviti po centraliziranem pristopu, temeljijo na podatkovno vodenem razvoju. Intervjuji z uporabniki zaradi pridobivanja zahtev niso zaželeni, saj so zahteve uporabnikov preveč variabilne. Prav to je najmočnejši protiargument dimenzijskemu modelu.

Nasprotno pa razvojni cikel pristopa gradnje porazdeljenega podatkovnega skladišča izhaja iz poslovnih zahtev. To je razumljivo, saj je v zgradbi porazdeljenega podatkovnega skladišča uporabljen dimenzijski model, ki je izpeljan iz zahtev. Zato je, nasprotno kot pri centraliziranem pristopu, zbiranje zahtev s pomočjo intervjujev zelo zaželeno.

Federativni pristop kot hibrid dopušča obe možnosti, čeprav močno zagovarja začetno zbiranje zahtev – tudi z intervjuji, ki vključuje natančne definicije in analizo poslovnih potreb. Federativna zgradba ima kot zgradba zgradb posebni razvojni cikel, ki se bistveno razlikuje od obeh ostalih.

| C | P | F |
|---|---|---|
| implementacija podatkovnega skladišča | definicije poslovnih zahtev | dokumentiranje obstoječih sistemov, izdelava E-R modela |
| ↓ | ↓ | ↓ |
| integracija podatkov | izdelava načrta in dimenzijskega modela | določitev kandidatov z dodano vrednostjo |
| ↓ | ↓ | ↓ |
| testiranje in izboljšave | specifikacija orodij za končnega uporabnika | analiza in izbira kandidatnih podatkov z največjo dodano vrednostjo |
| ↓ | ↓ | ↓ |
| programiranje | izbira in namestitev orodij, fizično oblikovanje | implementacija orodij za zajem, transformacijo in polnjenje |
| ↓ | ↓ | ↓ |
| oblikovanje sistema za podporo odločanju | načrt priprave podatkov in razvoj orodij | gradnja skladišča |
| ↓ | ↓ | ↓ |
| analiza rezultatov | razvoj aplikacij za uporabnika | dokumentiranje |
| ↓ | | ↓ |
| razumevanje zahtev | | uporaba |

Tabela 3: **Razvojni cikli zgradb (primerjalno)**

Prav tako lahko ugotovimo, da izbira podatkovnega modela ne vpliva na izbiro razvojnega cikla, temveč na to vpliva lastnost posameznega pristopa v celoti. Tako uporabljamo pri centraliziranem pristopu gradnje podatkovnega skladišča isti razvojni cikel za načrtovanje osrednjega podatkovnega skladišča kot tudi področnih skladišč, kljub njuni popolnoma različni vlogi, strukturi in namenu uporabe.

### 3.4 Drugi parametri

V kontekstu podatkovnega skladiščenja izluščimo še naslednje parametre pristopov, pri katerih se lahko posamezni pristopi razlikujejo:

- upravljanje meta podatkov (pri centraliziranemu pristopu so meta podatki nujno porazdeljeni);
- enostavnost začetne gradnje (trud, potreben za izgraditev podatkovnega skladišča ali prve delujoče področne shrambe);
- upravljanje (trud, potreben za učinkovito upravljanje "zakulisja");
- nadgradljivost (trud, potreben za prilagoditev na nove razmere, nastale zaradi novih zahtev uporabnikov ali organizacijskih sprememb);
- skalabilnost (stopnja možnih razširitev zaradi večjega števila uporabnikov, večje količine podatkov, večje podpore povzetim podatkom, bolj kompleksnih povpraševanj);
- varnost in zaščita;
- možnost vpeljave zunanjih sodelavcev;
- potrebna politična volja in
- razmerje med porabo virov na nivoju ZOS in posameznih oddelkov.

Omenjeni parametri se v praksi določijo nad končnimi orodji, kajti šele izbrani uporabljeni sistemi in orodja določajo končne lastnosti sistema podatkovnih skladišč.

V zaključku, v tabelah 4, 5 in 6, primerjalno zapišimo še najpomembnejše razlike med posameznimi pristopi:

| C | F |
|---|---|
| Monolitski sistem | Množica povezanih sistemov, ki si delijo skupne podatke |
| Atomarne informacije na enem, centralnem mestu | Atomarne informacije na različnih lokacijah |
| Homogeni sistem | Heterogeni sistem |

Tabela 4: **Centraliziran in federativni pristop (primerjalno)**

| P | F |
|---|---|
| Sestavljena je iz izključno področnih skladišč. | Mešane strukture (klasična centralizirana podatkovna skladišča kot tudi skladišča, sestavljena iz področnih skladišč) |
| Vsi podatki nastopajo na podatkovnem vodilu. | Ne zahteva nujno delitve vseh podatkov na podatkovnem vodilu. |
| Dimenzijski model je ekskluzivni model. | Priznava, da so atomarna podatkovna skladišča (E-R model) boljša rešitev v določenih okoljih, kjer je to dovoljeno in zaželeno. |

Tabela 5: **Centraliziran in porazdeljeni pristop (primerjalno)**

| C | P |
|---|---|
| Uporablja rigorozne, že znane tehnike za zbiranje, modeliranje in implementiranje zahtev končnih uporabnikov. | Dimenzijski model omogoča boljše razumevanje in boljšo učinkovitost v smislu hitrejšega izvajanja povpraševanj. |
| Temelji na področno usmerjenem podatkovnem modelu, ki minimizira integracijske probleme med posameznimi projekti podatkovnega skladiščenja. | Atomarne informacije na različnih lokacijah. |
| Dovoljuje gradnjo podrejenih, odvisnih področnih skladišč in s tem omogoča vodljivo uporabo tehnologije področnih skladišč. | Množica povezanih, vendar decentraliziranih področnih skladišč, ki si delijo skupne podatke. |
| Omogoča bolj centralističen nadzor nad sistemom podatkovnega skladišča. | Omogoča hitrejšo gradnjo sistema podatkovnega skladišča in njegovo hitrejše prilagajanje spremembam. |

Tabela 6: **Porazdeljeni in federativni pristop (primerjalno)**

### 3.5. Izbira primernega pristopa

Ne glede na izbrani pristop, strategijo in orodja mora uspešen projekt podatkovnega skladiščenja:

- zadostiti trenutnim zahtevam uporabnikov;
- biti upravljan s sprejemljivimi stroški;

- biti dovolj prilagodljiv glede na spremembe zahtev uporabnikov in organizacijske spremembe;
- nuditi konsistentne in visokokakovostne podatke in
- omogočati uporabnikom lažjo navigacijo in razumevanje podatkov ter jim pomagati pridobiti iz podatkov največ, kar se da.

Na podlagi izvedene primerjave v prejšnjem poglavju, v nadaljevanju predlagamo napotke za izbiro ustreznega pristopa. Strnjeno so predstavljeni v tabeli 7.

Naslednji dejavniki dajejo prednost uporabi centraliziranega pristopa pri gradnji podatkovnega skladišča:

- Stabilno, hierarhično strukturiran ZOS, kjer so usmeritve (razvojne, organizacijske) določene z višjega nivoja in niso predmet sodelovanja med posameznimi oddelki. Tako okolje poenostavlja integracijo, saj lahko v primeru neskladij in nesoglasij med oddelki učinkovito nastopi vodstvo.
- Velika želja po podpori odločanju, bodisi da je ta želja že prisotna med uporabniki, ali pa je to le zaveza vodstva.
- Stabilen in z viri močan oddelek informacijske tehnologije (IT) na nivoju ZOS.
- Oddelek IT na nivoju ZOS, ki dobro pozna poslovne probleme v organizaciji in ima sposobnosti in motivacijo za reševanje poslovno-nivojskih integracij, ki so del gradnje in vzdrževanja podatkovnega skladišča in ni le tehnično usmerjen.

Zadržki pri uporabi centraliziranega pristopa so naslednji:

- potrebna je precejšnja investicija finančnih sredstev in časa, zato je tudi tveganje večje;
- nujno je potrebno pokroviteljstvo s strani članov vodstva in to za celoten čas izvajanja projekta;
- za integracijo operativnih podatkov s celotnega ZOS v koherentno celoto je potrebno vključiti informatike z nivoja organizacije in ne oddelka, kar je prevečkrat nedosegljiv ideal, ker se informatiki soočajo z množico neintegriranih operativnih sistemov že na nivoju oddelkov in ne želijo ponavljati te izkušnje na projektu podatkovnega skladiščenja;
- težja prilagodljivost organizacijskim spremembam;
- daljša odzivnost na spremembo lokalnih zahtev in
- odvisnost področnega skladišča od osrednjega (v daljšem času).

Pristop gradnje porazdeljenega podatkovnega skladišča označujejo naslednje želene značilnosti:

- omogoča hitro gradnjo razmeroma enostavno obvladljivih področnih skladišč, ki predstavljajo rešitev prioritetnih (poslovnih) problemov;
- manjši projekti omogočajo preizkušanje metodologij, orodij in strategij, brez večjih izgub oziroma stroškov in usodnih vplivov na kasnejše dele projekta in ker
- omogoča razmeroma enostavno rešitev polnjenja iz operativnih sistemov v področno shrambo (znotraj posamezne iteracije in le za podatke na nivoju elementarnih transakcij).

Pri oblikovanju porazdeljene zgradbe podatkovnega skladišča je največja nevarnost v tem, da posamezna področna skladišča hitro postanejo nepovezana z ostalimi, torej "informacijski otoček sredi oceana". Osnovni razlog je v tem, da se želijo ali vodstvo ali projektni vodje izogniti zahtevani investiciji v začetne korake, ki med drugim definirajo skladne dimenzije in dejstva, skupna poslovna pravila in semantiko, kar rezultira v skupnih meta podatkih. Omenjeno je mogoče doseči le s trdim delom ob podpori vodstva. Nepovezljivost oziroma neintegriranost, ki je posledica množice samostojnih nepovezljivih področnih skladišč, krši enega izmed osnovnih razlogov za odločitev za podatkovno skladišče in predstavlja skoraj nepremostljivo oviro za nadaljnji razvoj podatkovnega skladišča na nivoju ZOS, saj je kasnejša integracija nemogoča ali vsaj zelo težavna.

Ostale pomanjkljivosti decentraliziranega pristopa:

- oddelki imajo svoje, njim lastne podatke, ki jih ne želijo deliti z drugimi oddelki;
- oddelki imajo svoje zahteve, zato mora podatkovno skladišče, sestavljeno le iz področnih shramb, optimalno integrirati prav vse zahteve vseh uporabnikov;
- težavna pogajanja o političnih (katera področna shramba je najpomembnejša in jo je potrebno prioritetno zgraditi) in tehnoloških odločitvah med oddelki;
- potreben dogovor o skupni zgradbi, poslovnih pravilih in semantiki med različnimi skupinami;
- validacija zgrajenega sistema v razvojnem ciklu ni eksplicitno podana;
- pristopa raje ne uporabimo, če imajo oddelki bistveno različne potrebe (manjša prekrivanost zahtev).

Federativna zgradba uspešno integrira množico komponent v sistemu: kupljena in zgrajena podatkovna in področna skladišča ter analitične aplikacije, podatkovno rudarjenje, orodja za sprotno analizo, orodja za povpraševanja in poročanje, orodja za izdelavo poročil iz operativnega dela, orodja za povečanje kakovosti podatkov, orodja za zajem, transformacijo in polnjenje podatkov, orodja za sistemsko upravljanje, orodja za dostavo informacij, informacijske duri ZOS, sisteme za poročanje in sistemi za upravljanje podatkovnih baz. Vsekakor velja pristop gradnje federativne zgradbe uporabiti pri množici sistemov podatkovnih skladišč znotraj ZOS, kar danes ni več redkost. V takem primeru lahko federativno zgradbo obravnavamo kot generično oziroma kot metamodel sistemov podatkovnih skladišč.

Za oblikovanje in razvoj federativne zgradbe se je potrebno odločiti ali vsaj razmišljati ob naslednjih situacijah:

- Ob prevzemu in prodaji podjetij – čemur smo dnevno priča tudi v Sloveniji – bi bilo neracionalno in nesmiselno zavreči popolnoma delujočo infrastrukturo podatkovnega skladišča, zato je za integracijo virov nujna prilagoditev na federativno zgradbo.
- Tržišče se premika v stopnjo razvoja, kjer se programski izdelki le še kupujejo in ne razvijajo več, in to velja tudi za podatkovna skladišča. Tako je lahko novi celoviti poslovni rešitvi (ERP, angl. Enterprise Resource Planning) ali sistemu OLTP priloženo tudi delno razvito podatkovno skladišče, ki ga je potrebno integrirati v že zgrajen sistem podatkovnega skladišča.
- V primeru več različnih sistemov podatkovnih oziroma področnih skladišč.
- V primeru analitičnih aplikacij, pogosto ne več vzdrževanih, ki jih je potrebno integrirati.

Pomanjkljivosti federativnega pristopa so:

- Razmeroma težavno usklajevanje in koordiniranje aktivnosti, potrebnih pri gradnji podatkovnega skladišča.
- Težavno "prebijanje ledu" pri političnih in lastniških odločitvah.
- Zahteva dogovor o poslovnih pravilih in semantiki med različnimi skupinami.
- Kompleksno in zahtevno tehniško okolje.
- Pogosto ima več repozitorijev metapodatkov.

Omenjene ugotovitve lahko strnemo v tabelo, ki predstavlja metodo za določitev optimalnega pristopa pri gradnji podatkovnega skladišča (tabela 7). Predstavljena je primernost posameznega pristopa glede na parameter, ki opisuje stanje v organizaciji ali določeno zahtevo, ki jo je potrebno zadovoljiti.

| Parameter \ Pristop | C | P | F |
|---|---|---|---|
| Nehierarhična organiziranost in nadzor ZOS | N | P | P |
| Potreba po hitri rešitvi | N | P | N |
| Potrebe po različnih natančnosti posameznih podatkov (različna granularnost) | P | N | N |
| Dinamične spremembe v organizaciji | N | P | P |
| Potreba po različnih virih (izvorih) podatkov | P | N | P |
| Z viri šibek oddelek IT na nivoju ZOS | N | P | P |
| Pokroviteljstvo projekta podatkovnega skladišča ni zagotovljeno | N | P | P |
| Želja po preizkušanju orodij | N | P | NP |
| Množica obstoječih rešitev v sistemu za analiziranje | N | N | P |
| Množica sistemov podatkovnih skladišč | PP | N | P |
| Možnost vpeljave zunanjih sodelavcev | P | PP | PP |
| Potreba po močnih varnostnih mehanizmih | P | N | N |
| Zahtevana odprtost modela | N | P | P |

Legenda: *P – primerna, N – neprimerna, PP – pogojno primerna*

Tabela 7: **Določitev optimalne zgradbe podatkovnega skladišča**

Splošno najboljši pristop ne obstaja, saj ima vsak svoje dobre in slabe lastnosti. Vsekakor težimo k optimalni zgradbi, ki je v večini primerov decentralizirana, porazdeljena in predstavlja najboljšo izbiro za večino organizacij. Za podrobnejše informacije je bralec napoten na [Golob 2001].

## 4 Povzetek ugotovitev in sklep

Naše ugotovitve ne potrjujejo ugotovitev v [Gallas 1999], kjer je postavljena hipoteza, da imata avtorja Inmon in Kimball razmeroma skladne pristope, le da se njuni poimenovanji struktur razlikujeta. Nasprotno, naše ugotovitve, zbrane v tem prispevku, nakazujejo na veliko razliko pri:

- pristopu h gradnji podatkovnega skladišča: različen razvojni cikel;

- zgradbi: od-zgoraj-navzdol (centralizirana) oziroma od-spodaj-navzgor (porazdeljena) in
- uporabi osnovnega podatkovnega modela: E-R oziroma dimenzijski model.

Ugotovitve prav tako potrjujejo neskladnost jezika, ki ga avtorja uporabljata za opis svojih metodologij (drugačno razumevanje vloge in lastnosti podatkovnega skladišča, področnih skladišč, operativne podatkovne shrambe).

Rezultati naše primerjave v tretjem poglavju prav tako kažejo nasprotno, kot trdi [Gallas 1999], ki navaja, da se oba eksperta strinjata v tem, da je uspeh podatkovnega/področnega skladišča najprej odvisen od učinkovitega zbiranja poslovnih zahtev, ki določajo nadaljnje oblikovanje skladišča. Pokazali smo namreč, da razvojni cikel centraliziranega pristopa postavlja analizo zahtev na zadnji korak, v inkrementalni pa je postavljen čisto na začetek.

Vendar na dokončno izbiro večkrat vplivajo zunanji dejavniki ali odločitve, na katere imamo le omejen vpliv ali pa ga sploh nimamo. V takem primeru so rezultati prispevka uporabni pri identifikaciji šibkih točk izbranega pristopa ali določene zgradbe.

Splošno najboljša rešitev ne obstaja, saj ima vsak pristop (in s tem zgradba) svoje lastnosti, dobre in slabe, potrebno pa se je odločiti za tisto, ki najbolj ustreza posamezni organizaciji.

## LITERATURA


Anahory, S., Murray, D. (1997):
"Data Warehousing in the Real World:
A practical Guide for Building Decision Support",
Systems Addison Wesley Longman Limited, Harlow, England.

Barquín, R.C., Edelstein, H. (1997):
"Planning and Designing the Data Warehouse", Prentice Hall, Upper Saddle River, N.J.

Bischoff, J., Alexander, T. (1997):
"Data Warehouse: Practical Advice from the Experts", Prentice Hall, Upper Saddle River, N.J.

Firestone, Joseph M. (1998):
"Dimensional Modeling and E-R Modeling In The Data Warehouse".

Gallas, S. (1999):
"Kimball Vs. Inmon.", DM Review.

Golob, I. (2001):
"Arhitekture podatkovnih skladišč", magistrsko delo, Univerza v Mariboru.

Hackney, D (2000a):
"Data Warehouse Delivery:
Federated FAQs.", DM Review.

Hackney, D. (2000b):
"Data Warehouse Delivery:
How to Federate", DM Review.

Hackney, D. (2000c):
"Data Warehouse Delivery:
The Federated Future", DM Review.

Hackney, D. (2000d):
"Data Warehouse Delivery:
When to Federate", DM Review.

Hammond, M. (2000):
ZDNet: eWEEK: Survey traces huge growth in data warehouse market. http://www.zdnet.com/

Inmon, W. H. (1996):
"Building the Data Warehouse", Wiley, New York.

Inmon, William H. (1999a):
"Data Mart Does Not Equal Data Warehouse", DM Review.

Inmon, William H (1999b):
"Operational Data Store".

Kimball, R. (1996):
"The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses", John Wiley & Sons, New York, Chichester.

Kimball, R. et al. (1998):
"The Data Warehouse Lifecycle Toolkit : Expert Methods for Designing, Developing and Deploying Data Warehouses", Wiley, New York.

Wells, D., Hope, M. (2000):
"Ovum Evaluates: the Datamart - the Sucessful Route to Data Warehousing", Ovum.

White, C. (2000):
"The Federated Data Warehouse", DM Review.



Izidor Golob je doktorski študent na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru, kjer je tudi zaposlen kot asistent za področje informatika. Na raziskovalnem področju se ukvarja s podatkovnimi bazami, kakovostjo informacij in podatkovnimi skladišči.

Tatjana Welzer je izredna profesorica na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru, kjer predava na dodiplomski in podiplomski stopnji in vodi laboratorij za podatkovne tehnologije. Na raziskovalnem področju se ukvarja predvsem s podatkovnimi bazami, kakovostjo podatkov in podatkovnim modeliranjem.

Boštjan Brumen je doktorski študent na Fakulteti za elektrotehniko, računalništvo in informatiko v Mariboru, kjer je zaposlen kot asistent za področje informatika. Na raziskovalnem področju se ukvarja s podatkovnimi bazami in podatkovnim rudarjenjem.