Podatkovna skladišča in kakovost podatkov

Krista Rizman Žalik

Povzetek

Podatkovna skladišča predstavljajo pomembno informacijsko podporo poslovanju in poslovnim odločitvam, saj hranijo podatke o poslovanju. Poslovne odločitve so natančne le, če so podatki točni, investicija v podatkovno skladišče pa bo poplačana samo, če so podatki zanesljivi. Kakovost informacij iz podatkovnega skladišča predstavlja resno tveganje, ki ga je treba obvladati ob načrtovanju in razvoju podatkovnega skladišča. V prispevku so predstavljene in analizirane metode za merjenje kakovosti podatkov podatkovnega skladišča in metode za zagotavljanje in povečanje kakovosti podatkov. Podana je kratka primerjava med njimi, iz nje pa izhajajo smernice za izbiro metode za zagotavljanje in izboljšanje kakovosti podatkov podatkovnega skladišča.

Abstract

Data Warehouses and Data Quality

Data warehouses collect data and provide an important information support to produce information for business decision making. It is successful only if complete and accurate data are applied. The investment will be returned only if data are reliable. The information quality provided by data warehouses is a serious risk, which should be taken into account during the process of data warehouse design and development. In this paper, methods for data quality measurement and methods for providing and increasing the data quality are analyzed. A short comparison of analyzed methods is given. Based on the comparison, directions for choosing a method to provide and improve data quality are given.

1 Uvod

Podjetja danes hranijo in sproti obdelujejo transakcije – poslovne dogodke v poslovnem procesu – v sistemih za sprotno obdelavo, ki jih označujemo tudi s kratico OLTP (angl. On-Line Transactional Processing). Za povečanje uspešnosti poslovanja pa gradijo podatkovna skladišča (angl. Data Warehouses), ki dajejo celovit pogled na podatke posameznega podjetja. Omogočajo izdelavo potrebnih analiz, opazovanj trendov in predvidevanj posameznih kazalcev poslovanja. Na voljo so analitikom in upravljalcem, ki potrebujejo informacije za odločanje v sprejemljivem času. Z uporabo analitičnih orodij (angl. OLAP – On-Line Analytical Processing) in podatkovnega rudarjenja predstavljajo podatkovna skladišča danes vrh informacijske podpore poslovanju, ki je stabilno in zanesljivo le, če so tudi podatki dovolj zanesljivi, točni in popolni. V nasprotnem primeru iz podatkovnega skladišča ne moremo dobiti zanesljivih informacij. Zato pri izgradnji podatkovnih skladišč ne smemo pozabiti na kakovost podatkov. Podatkom neznane kakovosti in rezultatom analiz, ki na takšnih podatkih temeljijo, ne moremo zaupati.

1.1 Kakšna je danes kakovost podatkov v podatkovnih skladiščih?

Kakovost podatkov v obstoječih podatkovnih skladiščih je pogosto slaba zaradi napak pri vnosu, različnih sprememb podatkov in struktur, ki so nastajale skozi čas, napačnih podatkov iz spletnih aplikacij ali iz zunanjih virov, ali pa zaradi združevanja dobrih podatkov z zastarelimi in nezmožnostjo ločevanja med njimi.

V poročilu leta 2002 [20], ki ga je izvedel *Data Warehousing Institute*, je skoraj polovica (44 %) vprašanih poročala, da je kakovost podatkov "slabša, kot si kdorkoli lahko predstavlja". 40 % anketiranih je priznalo, da so stroški, problemi in izgube neposredno povezani s slabo kakovostjo podatkov. *Data Warehouse Institute* v poročilu ugotaljavlja, da težave povezane s kakovostjo podatkov trenutno stanjejo ameriška podjetja 600 milijonov ameriških dolarjev na leto. V poročilu *Pricewaterhouse Coopers* [9] so ocenili tudi kakovost podatkov v podatkovnih skladiščih. Poročilo z analizo izjav vodilnih direktorjev iz 600 podjetij ugotavlja, da jih je 60 % zmanjšalo stroške poslovanja, več kot 40 % je povečalo prodajo z boljšimi analizami podatkov o strankah in več kot 30 % podjetij je podpisalo pomembne pogodbe zaradi boljše analize podatkov. To poročilo poudarja pomen kakovosti podatkov in povzema, da imajo podjetja, ki upravljajo svoje podatke kot strateški vir in investirajo v

njihovo kakovost, večji ugled in so bolj dobičkonosna kot tista, ki tega ne počnejo.

1.2 Prenova podatkovnih skladišč

Podatkovna skladišča danes hranijo podatke, ki največkrat odgovarjajo na splošna vprašanja: kateri izdelki prinesejo največji/najmanjši dobiček, kakšne izdelke in storitve bodo uporabniki želeli v prihodnje, kateri so sedaj in so bili v bližnji preteklosti najbolj prodajani izdelki. Značilni podatki, ki se zbirajo v podatkovnih skladiščih, so: enolična oznaka izdelka, izdelek, trgovina, cena, popust, količina, vrednost, datum. S temi ključnimi podatki lahko analiziramo prodajo po izdelkih, času, skladiščih in področjih. Ti običajni podatki podatkovnih skladišč pa ne povedo nič o kupcu, kaj vse je kupil in koliko, katere izdelke je kupil hkrati in v kakšnem časovnem zaporedju. Podjetja danes dodajajo informacije o svojih strankah v podatkovna skladišča, kar jim prinese konkurenčno prednost. 70–80 % podatkovnih skladišč svetu je namenjenih obravnavi strank. Podjetja težijo k pridobivanju novih strank - kupcev. Ker pa število le-teh ni neomejeno, so cilji podjetij tudi obdržati svoje stranke čim dlje, prodati jim čim več stvari in poslovati z njimi na učinkovitejši način. Da lahko te težnje udejanijo in povečujejo prodajo ter dobiček, si morajo ustvariti jasno sliko obnašanja svojih strank, pravila nakupov in analizirati njihove navade. Podatki o strankah, ki jih dodajajo v podatkovna skladišča, so manj formalizirani, strukturirani in zato tudi manj kakovostni v primerjavi s podatki, ki so že v podatkovnih skladiščih. Z dopolnjevanjem podatkovnih skladišč postane problem kakovosti podatkov za podjetja še večji.

2. Kaj je podatkovno skladišče in kaj kakovost podatkov?

Immon definira podatkovno skladišče kot subjektno orientirano, integrirano, stanovitno, časovno raznoliko zbirko podatkov, ki podpirajo poslovno odločanje [11]. Skladišče ni funkcionalno orientirano ampak so v njem podatki subjektov. V podatkovnih skladiščih CRM (angl. Customer Relationship Management) je subjekt stranka. Podatki so integrirani in poenoteni. Podatki so lahko elementarni ali že integrirani. Podatkovno skladišče hrani podatke stare tudi več let, kar omogoča analize trendov in oblikovanje napovedi gibanja posameznih kazalcev v prihodnje. Kimball [13,14] vidi podatkovno skladišče kot področne shrambe, kjer hranimo podrobne transakcijske podatke. V virih, kjer so primerjali obe definiciji, ne vidijo bistvene razlike med definicijami in poudarjajo, da se podatkovno skladiščenje razvija in dopolnjuje [1,7].

Obstaja več videnj kakovosti podatkov:

- *Current Analysis* [2] pravi, da je kakovost podatkov preprosto zrcalo točnosti podatkov organizacije. Dobra kakovost podatkov pomeni, da so podatki organizacije točni, popolni, konsistentni, pravočasni, enoviti in pravilni. Bolj kakovostni so podatki, bolj jasno predstavljajo natančen, usklajen vidik podjetja skozi podsisteme, organizacijske enote in vrste poslovanja.
- Kakovost podatkov je stanje popolnosti, veljavnosti, konsistentnosti, pravočasnosti in točnosti, kar dela podatke primerne in uporabne za specifične analize [10].
- Kakovost podatkov je često definirana kot proces urejanja informacij, tako da so posamezni zapisi natančni in točni, ažurni in konsistentno predstavljeni. Kakovost podatkov je težko izmeriti in je dovolj kritičen dejavnik za neuspeh projekta ali neizpolnitev strateškega cilja [3,5,6]. Podatkovna plast organizacije je kritični element, ker je zelo enostavno zanemariti kakovost podatkov ali narediti preoptimistične ocene. Imeti kakovostne podatke za analizo je velika konkurenčna prednost.

3. Kako merimo kakovost podatkov?

Že iz širine zgornjih definicij o kakovosti podatkov je jasno, da bomo poskušali le ovrednotiti kakovost podatkov, ne bomo pa je mogli natančno izmeriti. Meritev kakovosti podatkov ni delo za perfekcioniste, ampak za ljudi, ki poznajo pomen meritve. Ogledali si bomo dve metodi za merjenje kakovosti podatkov. Prva, opisana v viru [15], zelo poudarja pomen množice osnovnih pravil za ocenitev kakovosti vhodnih podatkov, ki so pomembni za uspešnost gradnje podatkovnega skladišča glede na vsebino in pravočasnost.

Nekaj primerov pravil:

- samo izdelki dodani v zadnjem mesecu nimajo natančnih informacij o prodaji,
- vsi izdelki imajo enolične identifikatorje v tabeli prodaj izdelkov,
- ni presledkov pred imeni kupcev,
- ni posebnih znakov pri imenih kupcev,
- vsi identifikatorji izdelkov so enolični,
- vse prodaje so med 0 in 100.000.000,00 SIT itn.

Ko so pravila definirana in so določeni podatki, moremo oceniti kakovost podatkov. Ocenitev bo predstavljala stanje kakovosti podatkov za to pravilo. Celotna ocena bo povprečje ocen izraženo v točkah. Samo povprečje ni tako učinkovito kot sistem, ki da večjo težo pomembnejšim pravilom. Vrednost celotne ocene (število točk) definira kakovost podatkov. Avtor iz empiričnih izkušenj predlaga ocenitev, ki da naslednje rezultate prikazane v tabeli 1 [15].

Ocena kakovosti	
99–100	izjemna kakovost podatkov
96–98	dobra kakovost podatkov
90–95	povprečna kakovost podatkov
81–89	podpovprečna kakovost podatkovTakšna kakovost negativno vpliva na poslovanje.
65–80	slaba kakovost podatkovSistem bo največkrat dal zelo slabe rezultate.
0–64	zelo slaba kakovost podatkovPodatki so neuporabni in jih treba izboljšati.

Tabela 1: **Ocenitve kakovosti podatkov**

Druga metoda, opisana v viru [10], opravi meritev kakovosti podatkov z opazovanjem vseh lastnosti kakovosti podatkov. V tabeli 2 so primeri meril, katerih kršenje nam v odstotkih predstavlja nekakovost podatkov. Tabela prikazuje množico skupnih lastnosti za merjenje kakovosti podatkov in navaja primere meril za določanje zahtev za kakovost.

3.1 Primerjava metod merjenja kakovosti

Obe obravnavani metodi merjenja kakovosti podatkov temeljita na postavljanju pravil in nato na ocenitvi odstotka primerov podatkov, ki kršijo postavljena pravila in predstavljajo odstotek nekakovostnih podatkov.

Druga metoda oceni kakovost podatkov bolj natančno, z vseh vidikov definicij kvalitete podatkov podatkovnega skladišča. Najde nekakovostne podatke, ki kršijo pravila dobre kakovosti podatkov, in zahteva njihovo zmanjšanje. Ni pa merila, ki bi definiralo, kaj pomenijo odstotki nekakovosti podatkov. Ocena ukrepanja je v rokah razvijalca podatkovnega skladišča.

Prva metoda daje empirično merilo za opisno oceno kakovosti podatkov iz povprečja odstotkov neujemanja pravil. Metoda poudarja določitev množice osnovnih pravil za ocenitve kakovosti vhodnih podatkov, ki so pomembni za uspešnost in dobičkonosnost gradnje podatkovnega skladišča glede na vsebinske in časovne omejitve.

Meritev kakovosti podatkov lahko izvedemo z orodji (na primer Data Quality Inspector [17]), ki omogočajo definiranje pravil in izbiro podatkovnih množic. Nato analizirajo podatke in poiščejo kršitve pravil ter posredujejo nekakovostne zapise odgovornim za njihovo kakovost.

4 Možnosti povečanja kakovosti podatkov

Obstajajo tri možnosti povečanja kakovosti podatkov:

1. Odložiti aktivnost povečanja kakovosti podatkov na kasneje.

Latnosti kakovosti podatkov	**Opis**	**Primer merila**
Natančnost	Natančnost je razmerje ujemanja med množico podatkov in pravilnih vrednosti.	Merilo natančnosti je odstotek pravilnih vrednosti.
Popolnost	Popolnost je razmerje, ki opisuje število atributov, ki imajo določene vrednosti.	Merilo popolnosti je odstotek atributov, ki imajo vrednosti.
Doslednost	Doslednost je ujemanje ali logična skladnost med podatki.	Merilo doslednosti je odstotek ujemanja vrednosti, ki zadovoljujejo pogoje ujemanja ali skladne vrednosti.
Primerljivost	Primerljivost je ujemanje ali logična usklajenost, ki dovoljuje korelacije in primerjave z ostalimi podobnimi podatki.	Merilo primerljivosti je odstotek referenčne integritete.
Pravočasnost	Pravočasnost je dostopnost do podatka ali več podatkov, ki je zagotovljena v zahtevanem ali določenem času.	Merilo pravočasnosti je odstotek podatkov zagotovljenih v predpisanem času (na primer: v eni uri ali enem dnevu).
Enoličnost	Enoličnost je pojav, ki opisuje, da se vrednost atributa pojavi le enkrat.	Merilo enoličnosti je odstotek zapisov, ki ne ustrezajo principu enoličnosti.
Veljavnost	Veljavnost je lastnost, ki opisuje ustreznost in spremenljivost podatkov in zmanjšuje verjetnost napak.	Merilo veljavnosti je odstotek vrednosti, ki sodijo v pričakovane zaloge dovoljenih vrednosti.

Tabela 2: **Množica lastnosti kakovosti podatkov**

2. Vgraditi pravila, ki bodo preverjala podatke v aplikacijah, ki zbirajo podatke.
3. Zgraditi podatkovno skladišče in čistiti podatke.

Odločitev za eno izmed možnosti za povečanje kakovosti podatkov je odvisna od:

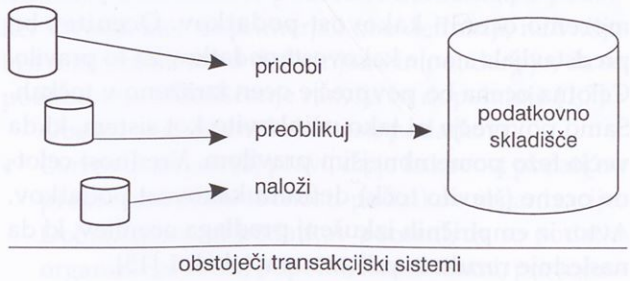
- stopnje kakovosti podatkov,
- stanja obstoječih transakcijskih sistemov (ali so že potrebni prenove ali pa so to novejši sistemi) in
- potreb in želja po poslovanju s kakovostnimi podatki.

Prva možnost je nasprotje reka »Kar lahko storiš danes, ne odlašaj na jutri« in se v praksi največkrat slabo obnese.

Druga možnost je nujno potrebna, če so podatki zelo slabi, tako rekoč neuporabni. Vgradnja pravil v aplikacijah, ki zbirajo podatke, pa ne poveča kakovosti starih podatkov. Le-ti pa so v podatkovnih skladiščih še kako potrebni za izvajanje analiz in primerjanj v daljšem časovnem obdobju in za napovedovanje gibanj posameznih količin v prihodnje. To možnost izvedejo v podjetjih, kjer so transakcijski sistemi zastareli in so potrebni prenove.

Tretja možnost je izboljšanje kakovosti podatkov v podatkovnem skladišču. Nalagati in hraniti nekakovostne podatke in jih nato kasneje prečistiti, je veliko večji strošek, kot pa prečistiti podatke in jih čiste hraniti v podatkovnem skladišču.

Mesto za zagotavljanje kakovosti podatkov v okolju podatkovnega skladišča je sam razvoj le-tega. Podatkovno skladišče s še tako izpopolnjenim načrtovanjem, a s podatki, ki niso kakovostni, je malo vredno. Zgrajeno podatkovno skladišče za upravljanje podatkov podatke tudi očisti, kar je pomembna naloga, ki je bila poudarjena že ob definiciji podatkovnega skladišča. Immon, oče podatkovnih skladišč, pravi, da je naloga koraka ETL (angl. Extract Transform Load, slov. pridobivanje, preoblikovanje in nalaganje podatkov) naložiti v podatkovno skladišče integrirane in prečiščene podatke (slika 1). Usklajenost procesa ETL in kakovost podatkov daje možnost za enostavnejšo upravljanje kompleksnih podatkovnih integracij. Z uvedbo kakovosti podatkov v proces ETL sta zagotovljena kakovost podatkov in točnost. Najtežavnejše v procesu ETL je pridobivanje podatkov iz različnih virov, preoblikovanje podatkov v nove formate in nalaganje podatkov v podatkovna skladišča. Cilj procesa ETL mora biti tudi zajemanje čistih in točnih podatkov.



Slika 1: **Proces ETL**

Za izboljšanje kakovosti podatkov v procesu ETL uporabljamo naslednje tehnike: čiščenje podatkov, dopolnjevanje in ujemanje ter usklajevanje.

Čiščenje podatkov je odkrivanje in popravljanje nekakovostnih podatkovnih elementov in podatkovnih struktur.

S tehniko dopolnjevanja povečujemo obseg informacij, ki jih lahko dobimo iz podatkov. Na primer, naslov lahko dopolnimo z vljudnostnimi nazivi, ki so odvisni od spola ali pa z geolokacijskimi podatki o pošti in mestu.

Z ujemanjem ugotovimo množico ponavljajočih se ujemajočih zapisov. S postopkom združevanja takšne zapise uskladimo in združimo v en zapis. Poznamo tri metode ujemanja in združevanja: odpravo podvojevanj, združevanje v družine in povezovanje zapisov. Vsaka izmed njih je uporabna v določenem primeru. Odprava podvojevanj z ugotavljanjem ujemanja in združevanjem, odpravi podvojene zapise. Združevanje v družine združi zapise, ki imajo vsaj en enak atribut. Ta atribut je ključ za združevanje. Tipičen primer je združevanje podatkov o strankah v skupna gospodinjstva po skupnem naslovu. Povezovanje zapisov je splošnejši primer združevanja v družine. Zapise povezujemo z drugim namenom in ne po skupnem naslovu, ampak na primer po organizacijskem principu.

Poznamo mehko, popolno in verjetnostno ujemanje. Mehko ujemanje temelji na bolj ali manj natančnih pravilih za ujemanje in na območjih podatkov, ki niso natančno definirana. Popolno ujemanje daje enako težo različnim atributom zapisa. Verjetnostno ujemanje izkorišča statistično verjetnost, da ujemanje določenega elementa zapisa z določeno verjetnostjo zagotavlja, da sta zapisa ista.

Čiščenje, dopolnjevanje, ujemanje in združevanje izvajamo zaporedno. Izvedemo samo tiste tehnike, ki so primerne in ustrezne našim problemom in podatkom.

Graditelji podatkovnih skladišč, kot je na primer Oracle Warehause Builder [19], so orodja, ki nudijo podporo razvoju in avtomatizirajo obravnavane tehnike povečanja kakovosti podatkovnih skladišč. Poleg razvojnih orodij obstajajo tudi razvojne metode, ki poudarjajo povečanje kakovosti [4,10,12,15,18]. Obravnavali in primerjali bomo:

- pristop k povečanju kakovosti podatkov zasnovan na tveganjih,
- pristop k povečanju kakovosti podatkov in ROI,
- metodo za razvoj podatkovnih skladišč DWM.

5 Pristop k povečanju kakovosti podatkov zasnovan na tveganjih

Ta vključuje štiri ključne aktivnosti za zagotavljanje kakovosti podatkov v razvojni cikel [10]:

1. Določitev pričakovane kakovosti podatkov in metrike za merjenje.
2. Identificiranje tveganja v kakovosti podatkov in predvidevanje, kdaj in kateri podatki iz podatkovnega skladišča ne bodo uspeli zadovoljiti pričakovanj.
3. Zmanjšanje tveganja – določitev akcije za zmanjšanje vsakega večjega tveganja.
4. Opazovanje in ocenitev rezultatov.

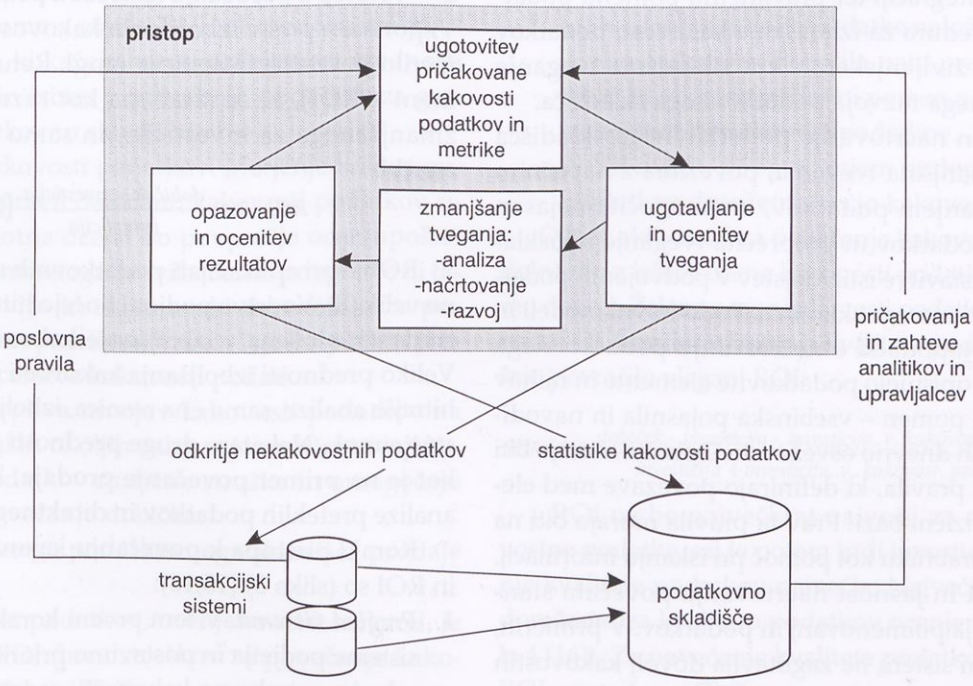
5.1 Določitev pričakovane kakovosti podatkov in metrike za merjenje

Prvi korak pri zagotavljanju kakovosti podatkov so definicije kakovosti podatkov, ki bodo pomagale uresničiti poslovne cilje. V primeru poslovnega cilja povečanja trga, so pričakovanja za kakovost podatkov osredotočena predvsem na zbiranje podatkov o prodaji strankam in kakovosti agregiranih podatkov.

Skupne lastnosti kakovosti podatkov po definirani metriki (tabela 2) določimo predvsem za podatke, ki povzročajo večje tveganje pri doseganju zastavljenega poslovnega cilja. Tveganje lahko povežemo s ciljem podatkovnega skladišča. Na primer, za doseganje poslovnega cilja optimizacije dobave so zelo pomembni podatki o razdeljevanju izdelkov.

5.2 Identificiranje tveganja v kakovosti podatkov

Je postopek predvidevanja, zakaj ne bi podatki iz podatkovnega skladišča uspeli zadovoljiti pričakovanj. Tveganje je lahko zunanji dogodek ali poznano stanje, ki povzroči nekakovostne podatke. Izpad računalnika je tveganje za pravočasnost podatkov. Na



Slika 2: **Zagotavljanje kakovosti podatkov po pristopu, zasnovanem na tveganjih**

primer, podatki o velikosti dobav izdelkov so obvezni podatki, toda v primeru izpada sistema kar za nekaj zapisov teh podatkov ne bo. Veliko tveganje je tudi napačna uporaba podatkov. Uporabniki podatkovnega skladišča v prvem koraku preverijo podatke, ki jih dobijo iz podatkovnega skladišča glede na podatke virov in pomagajo oblikovati ustrezna navodila za pravilno uporabo podatkov. Seveda pa v procesu gradnje sami preverimo podatke in ocenimo njihovo kakovost v skladu z merili, postavljenimi v prejšnjem koraku. Za to lahko uporabimo orodja, kot je na primer Data Quality Inspector [17]. Če ta preverjanja pokažejo, da je kakovost podatkov zajeta v podatkovno skladišče tako majhna, da predstavlja tveganje za dosego pričakovanih rezultatov podatkovnega skladišča, potem moramo oceniti napor za zmanjšanje tveganj in izvesti aktivnosti, ki zagotovijo večjo kakovost vhodnih podatkov.

5.3 Zmanjšanje tveganja

Za vsako tveganje definiramo kako in v katerem koraku razvoja ga bomo odpravili.

Procedura za izboljšanje kakovosti podatkov v različnih korakih razvoja podatkovnega skladišča je naslednja:

- V koraku definiranja zahtev je poudarek na metapodatkih, ki so pomembni za podporo analizi domen in integraciji ter pravilnemu pomenu podatkov. Procedura za izboljšanje kakovosti podatkov zgodaj v življenjskem ciklu zmanjšuje tveganje neuspešnega razvoja podatkovnega skladišča.
- Analiza in načrtovanje podatkovnega skladišča lahko zmanjšata tveganja, povezana z napačnim razumevanjem podatkov, onemogočita nejasen pomen podatkov in preprečita tveganje neusklajene predstavitve istih dejstev v podvojenih podatkih podatkovnega skladišča. Podatkovni modeli in ostali metapodatki o načrtovanju podatkovnega skladišča opisujejo podatkovne elemente in njihov natančen pomen – vsebinska pojasnila in navodila, kako jih dnevno osveževati. Opisana morajo biti poslovna pravila, ki definirajo povezave med elementi v fizični bazi. Prav ta pravila morajo biti na voljo uporabniku kot pomoč pri iskanju informacij.
- Izraznost in jasnost načrtovanja povečata standardizacija poimenovanj in podatkov. V primerih, ko izvorni sistem ne zagotavlja dovolj kakovostih virov, je smiselno začeti projekt za izboljšanje izvornega sistema, tako da bo zmanjšano tveganje zaradi nekakovostih podatkov. Iz zahtev za kakovost podatkov razvijemo množico pravil in metrike kakovosti, načrtamo in razvijemo preverjanja za pravila kakovosti podatkov za ugotavljanje in poročanje o napačnih podatkih. Onemogočimo vstop nekakovostnih podatkov v podatkovno skladišče.
- Razvoj lahko zmanjša tveganje z vključevanjem popravljanj in merjenj kakovosti podatkov v proces ETL. S preverjanjem ugotovimo preveri format in tip podatkov, usklajenost vrednosti z domenami, usklajenost z drugimi povezanimi podatki in usklajenost z metapodatki. Ugotavljanje pravilnosti preverja točnost podatkov, torej natančen opis realnih objektov. Potrjevanje pravilnosti pokaže primernost uporabe podatkov za različne uporabe in zmanjša tveganje napravilne uporabe podatkov.

5.4 Vzdrževaje podatkovnega skladišča – opazovanje, pregledovanje in ocena kakovosti podatkov

Kakovost vhodnih podatkov v podatkovno skladišče je potrebno nadzirati. Večkratna merjenja kakovosti vhodnih podatkov nam kažejo, ali se res približujemo zastavljeni kakovosti podatkov. Vzdrževanje pa predstavlja tudi ocenitev novih uporab podatkov in ugotavljanje vzrokov slabe kakovosti podatkov in njihovo odpravljanje.

6 Pristop k povečanju kakovosti podatkov in ROI

Vzporeden pojav izboljšanja kakovosti podatkov je merilo povračila investicije (angl. Return On Investment – ROI), ki se izračuna kot razmerje dobička zmanjšanega za investicijo, in samo investicijo po enačbi:

$$ROI = 100\,\frac{dobiček - investicija}{investicija}\,[\%]$$

ROI je pri aplikacijah podatkovnih skladišč običajno velik [8]. Vodstva podjetij hočejo hitro povračilo za sredstva, vložena v izboljšanje kakovosti podatkov. Veliko prednosti izboljšanja kakovosti je nemerljivih: hitrejše analize, samo ena resnica, izboljšano zadovoljstvo strank. Nekatere druge prednosti pa so merljive, kot je na primer povečanje prodaje, ki je posledica analize preteklih podatkov in direktnega oglaševanja.

Koraki pristopa k povečanju kakovosti podatkov in ROI so (slika 3) [15,16]:

1. *Pregled sistema.* V tem prvem koraku pregledamo sisteme podjetja in postavimo prioritete glede uporabe in potrebe po kakovosti podatkov. Ti sistemi temeljijo na točnosti podatkov in jih lahko najdemo

med sistemom OLTP, odločitvenimi sistemi in podatkovnimi skladišči.

2. Definiranje pravil. Pravil za kakovost podatkov ne dobimo z določitvijo, kako naj bodo podatki videti, ampak z oceno škode sistema, ki jo povzročijo nekakovosti podatki. Meritev vseh pravil daje oceno kakovosti podatkov. Pravilom, ki opisujejo zahteve za kakovost podatkov, ki, če so nekakovostni, povzročijo največ škode, damo največji pomen in težo. Nabor elementov kakovosti, ki jih lahko zagotavljajo pravila, je:
   - Referenčna integriteta, ki se nanaša na integriteto sklicevanja med podatki v različnih tabelah. Kot primer, identifikatorji izdelkov v tabeli prodaje izdelkov morajo biti natančneje opisani v tabeli izdelkov.
   - Enoličnost nekaterih podatkov, kot so identifikatorji. Primer kršitve pravila je, če isti identifikator opisuje dva ali več izdelkov.
   - Kardinalnost povezav določa količinska razmerja povezav. Izdelek z enim enoličnim identifikatorjem lahko prodamo večkrat ali pa ga še nismo nikoli prodali, če gre za nov izdelek.
   - Smiselne vrednosti domen.
   - Oblikovne pravilnosti, kot npr. nobenega presledka pred imeni, imena brez posebnih znakov, uporaba velikih in malih črk in ne samo velikih.
3. Označitev podatkov. Ugotavljanje lastnosti podatkov izvedemo s stavki SQL ali podobnimi povpraševanji, ki kažejo na porazdeljenost vrednosti podatkov in izbiro pomembnih podatkov za grajeno podatkovno skladišče.
4. Meritev kakovosti podatkov. Merjenje kršitev posameznih pravil da oceno kakovosti podatkov sistema. Celotna ocena bo povprečje ocen upoštevanja posameznega pravila. Samo povprečje ni tako učinkovito kot sistem, ki da večjo težo pomembnejšim pravilom. Tabela 2 je empirična preglednica ocenitve kakovosti podatkov.
5. Ocenitev vpliva nivojev kakovosti podatkov na ROI. V enakih okoljih zahtevajo različni nivoji kakovosti podatkov različne stroške in omogočajo različno povečanje dobička.

Primer iz realnega podatkovnega skladišča prodaje kaže, [16]:

- Za kakovost podatkov, ocenjeno z 90 točkami, je projektni ROI (brez stroškov za povečanje kakovosti podatkov) 175 %.
- Za kakovost podatkov, ocenjeno s 85 točkami, je projektni ROI (brez stroškov za povečanje kakovosti podatkov) 101 %.
- Za kakovost podatkov, ocenjeno z 80 točkami, je projektni ROI (brez stroškov za povečanje kakovosti podatkov) 64 %.

Tako vidimo, da je investicija v podatkovno skladišče (brez upoštevanja stroškov povečanja kakovosti podatkov) tem bolje poplačana, čim bolj kakovostne podatke obdelujemo. ROI (brez upoštevanja stroškov povečanja kakovosti podatkov) se nelinearno povečuje in je večji od linearnega prirastka za najbolj kakovostne podatke.

6. Povečanje kakovosti podatkov: Velja, da je ceneje povečati kakovost podatkov v virih kot pa v podatkovnih skladiščih. Če imamo na primer polje, ki ima veliko analitično vrednost in je polovica vrednosti nevnešenih in neuporabnih, potem nobena transformacijska strategija ne more pridobiti zanesljivih informacij iz teh podatkov. V primerih zastarelih sistemov povečanje v virih ni možno zaradi kompleksnosti in stare tehnologije. Kakovost podatkov povečamo z naslednjo množico aktivnosti v procesu ETL:
   - Podatke, ki zelo kršijo pravila kakovosti podatkov, naložimo v vmesne tabele, kjer jih lahko ročno pregledujemo in popravljamo.
   - Kršitve pravil kakovosti podatkov ugotovimo, o njih poročamo, toda podatke naložimo v podatkovno skladišče. Poročila nato pregledajo analitiki, ki poznajo poslovni pomen podatkov in so zadolženi za kakovost podatkov.
   - S čiščenjem, popravljanjem in dopolnjevanjem podatkov dosežemo večjo kakovost podatkov.

Obseg aktivnosti za povečanje kakovosti podatkov določimo z izbiro ocene kakovosti podatkov, ki jo želimo doseči. Želena ocena kakovosti podatkov je povezana z ROI. Stroški povečanja kvalitete podatkov zmanjšajo povračilo vlaganj ROI:

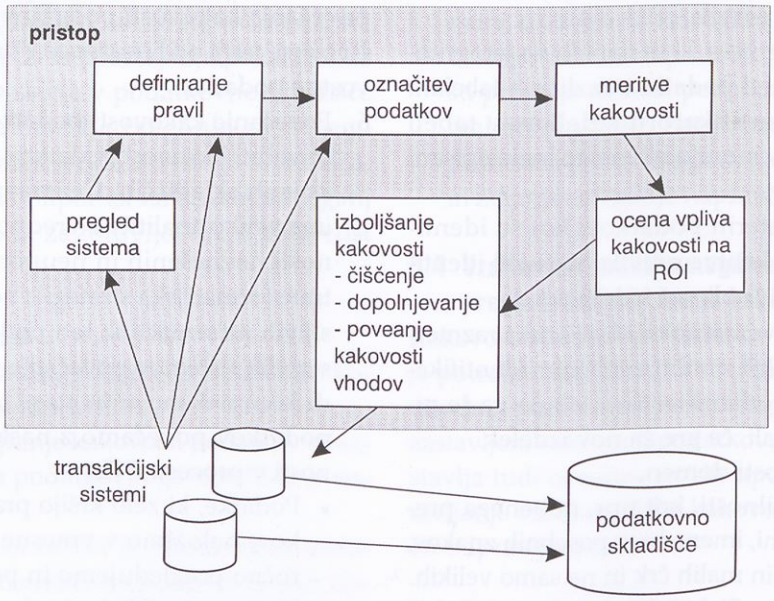
$$ROI = 100\,\frac{dobiček - investicija - investicija..v..kakovost..podatkov}{investicija + investicija..v..kakovost..podatkov} - [\%]$$

ROI ne bo največkrat največji za najbolj kakovostne podatke, saj je potem tudi investicija v povečanje kvalitete podatkov največja. Največji ROI (83 %) dosežemo za kvaliteto podatkov ocenjeno z 90 (tabela 4 [16]). Za povečanje kvalitete podatkov do 99 % se ROI zmanjša na 72 %.

Ocena kvalitete podatkov	Direktno trženje (A)	Uspešnost trženja (B)	Povprečen dobiček ($) (C)	Dobiček ($) (D=A*B*C)	Investicija* (I)	ROI* (=D-I/I)	Stroški povečanja kvalitete (I1)	ROI (=D-I-I1/I+I1)
99	115.000	0,0425	250	1221875	410000	198 %	410000	72 %
90	110.000	0,04	250	1100000	400000	175 %	400000	83 %
85	100.000	0,03	250	787500	390000	102 %	390000	69 %

** brez stroškov za povečanje kakovosti podatkov*

Tabela 4: **Stroški in koristi povečevanja kvalitete podatkov**



Slika 3: **Koraki pristopa k povečanju kakovosti podatkov in ROI**

7 Razvojna metoda DWM in povečevanje kakovosti podatkov

Metodologije za razvoj podatkovnih skladišč natančno opisujejo vse aktivnosti in procese v razvoju podatkovnega skladišča ter tudi bolj ali manj uspešno obravnavajo zagotavljanje kakovosti podatkov. Oglejmo si aktivnosti, ki jih poudarja metoda DWM (angl. Data Warehouse Method) [18] za zagotavljanje in povečanje kakovosti podatkov.

V prvem koraku razvoja podatkovnega skladišča (strategiji) je potrebno postaviti tudi strategijo kakovosti podatkov: pristop za sprotno integriteto podatkov v podatkovnem skladišču vključno z upravljanjem podatkov, upravljanjem z napakami in izjemami, čiščenjem podatkov, opazovanjem in nadzorovanjem podatkov. Metoda predlaga izdelavo lastništva podatkov in procese odprave odstopanj podatkov ter definiranje standardov za podatke.

Kakovost podatkov in integriteta morata biti obravnavani že v strategiji. Vpliv kakovosti podatkov na rezultirajoče podatkovno skladišče je zelo pomemben, saj kakovost podatkov določa zaupanje uporabnikov v podatkovno skladišče. Pri tem skrbimo za kakovost podatkov v vseh korakih razvoja.

- Pri definicijah določimo obseg zagotavljanja kakovosti podatkov ter izdelamo načrt. Definiramo naloge, vire in časovne okvirje za zagotavljanje kakovosti podatkov.
- Pri analizi izdelamo plane in postopke za povečanje kakovosti podatkov. Izdelamo natančne zahteve za kakovost podatkov vključno z obravnavanjem napak in izjem, čiščenjem podatkov in opazovanjem ter nadzorovanjem podatkov. Ustvarimo procedure za sprotno reševanje nekakovostnih podatkov (uporabe podatkov, standardov poimenovanja, formati, definicije podatkov, lastništva, popravljanje

napak). Metapodatki opisujejo tudi kakovost podatkov.

- V koraku načrtovanja načrtamo module, ki bodo izvedli zahtevane aktivnosti za zagotavljanje kakovosti podatkov, module za čiščenje, za obravnavo napak in izjem ter module za opazovanje in nadzorovanje podatkov.

Tabela 3 kaže prisotnost aktivnosti zagotavljanja kakovosti podatkov v vseh korakih razvoja podatkovnega skladišča in odstotek, potreben za proces zagotavljanja kakovosti podatkov v posameznem koraku razvoja [11].

PROCESI / FAZE	Zbiranje zahtev	Kakovost podatkov	...	Skupaj %
Strategija	...	2,2 %	...	100 %
Definicije	...	12,1 %	...	100 %
Analiza	...	6,5 %	...	100 %
Načrtovanje	...	7,9 %	...	100 %
Izgradnja	...	0,6 %	...	100 %
Uvedba	...		...	100 %

Tabela 3: **Kakovost podatkov v DWM je ena izmed aktivnosti, ki pokriva večino korakov načrtovanja**

8 Primerjava metod

Identificirali bomo posamezne elemente metod in pristopov ter izvedli primerjalno analizo med njimi. Za določitev učinkovitega pristopa pri zagotavljanju kakovosti podatkov podatkovnega skladišča, ki bo učinkovito izrabljal vire in nudil najobsežnejše možne analize iz obstoječih podatkov, je potrebno poznavanje lastnosti posameznih pristopov. Primerjali smo:

- izhodišča,
- način merjenja kakovosti,
- razvojni cikel,
- druge parametre (pomen metapodatkov, potrebna pripravljenost v večja vlaganja v podatkovna skladišča).

8.1 Izhodišča

Metode za razvoj podatkovnih skladišč in tudi obravnavana razvojna metoda DWM, obravnavajo kakovost podatkov kot pomembno in jo vključujejo v vse korake razvoja podatkovnega skladišča. Zagotavljanje kakovosti podatkov obravnavajo kot pogoj za uspešno izvedbo projekta.

Pristop k povečanju kakovosti podatkov zasnovanem na tveganjih gleda na proces upravljanja podatkov z vidika odprave tveganj za učinkovito zagotavljanje informacij iz podatkovnega skladišča.

V pristopu k povečanju kakovosti podatkov in ROI je izhodišče izboljšati kakovost podatkov tako, da bo učinek podatkovnega skladišča in ROI največji.

8.2 Način merjenja kakovosti

Oba primerjana pristopa predlagata za merjenje definiranje pravil in meritev odstopanj podatkov od definiranih pravil. V pristopu k povečanju kakovosti podatkov in ROI določimo pravila za kakovost podatkov na osnovi škode za učinkovito delovanje sistema; večja je le-ta, pomembnejše je pravilo. Pravil za kakovost podatkov ne dobimo z določitvijo, kakšni naj bodo podatki, ampak z oceno vpliva nekakovostnih podatkov na uspešnost delovanja sistema. Pravilom, ki opisujejo zahteve za kakovost podatkov, ki, če so nekakovostni, povzročijo največjo škodo, damo največji pomen in težo.

V pristopu k povečanju kakovosti podatkov zasnovanem na tveganjih definiramo pravila glede na skupne lastnosti kakovosti podatkov po definirani metriki (tabela 1) predvsem za podatke, ki povzročajo večje tveganje.

8.3 Razvojni cikel

Ugotovimo lahko, da pa se predstavljeni razvojni cikli obravnavanih metod bistveno ne razlikujejo med sebo (tabela 4). Pristop za povečanje kakovosti podatkov zasnovan na tveganjih, veliko bolj natančno opisuje in poudarja, kako v posameznem koraku razvoja pazimo na kakovost, kot ostale obravnavane metode. To je v skladu z izhodiščem pristopa. Pristop k zagotavljanju kakovosti in ROI pa daje večji poudarek ocenitvi vpliva nivojev kakovosti in stroškom izboljšav kakovosti podatkov, saj izhaja iz drugega izhodišča (tabela 4).

8.4 Drugi parametri

Pristop k povečanju kakovosti podatkov zasnovanem na tveganjih daje večji pomen metapodatkom kot pristop k povečanju kakovosti podatkov in ROI, saj natančno opisani metapodatki zmanjšujejo tveganja razvoja in napačnega razumevanja in uporabe dobljenih podatkov iz podatkovnega skladišča. Tudi razvojna metoda DWM daje velik poudarek metapodatkom, saj niso pomebni le za kakovost ampak tudi za uspešen razvoj.

Pristop k povečanju kakovosti zasnovanem na tveganju	Pristop k povečanju kakovosti in ROI	Razvojna metoda DWM
Določitev kakovosti (pravil)	Pregled sistema	Strategija (tudi kakovosti)
↓	↓	↓
	Definiranje pravil	Definicije
	↓	↓
Identifikacija tveganja	Označitev podatkov Meritev kakovosti	Analiza
↓	↓	↓
	Ocenitev vpliva nivojev kakovosti na ROI	
Zmanjšanje tveganja	Izboljšanje kakovosti	Izgradnja
↓		↓
Opazovanje in ocenitev rezultatov		Uvajanje in vzdrževanje

Tabela 4: **Razvojni cikli**

Za pristop zasnovan na tveganjih je potrebna pripravljenost podjetja v večja vlaganja v podatkovna skladišča, saj je bolj celovit in dosleden in sam ne poudarja dobičkonosnosti investicije v podatkovno skladišče, tako kot pristop k povečanju kakovosti podatkov in ROI.

8.5 Izbira pristopa povečanja kakovosti podatkov

Pomembna razlika med pristopi je v pojmovanju vloge podatkovnega skladišča. Pri prvem obravnavanem pristopu k zagotavljanju kakovosti podatkov zasnovanem na tveganjih, mora biti podatkovno skladišče brezhiben in izjemno kakovosten vir ter zelo zanesljiva informacijska podpora za poslovno odločanje. Pri drugem obravnavanem pristopu, kjer gre za povečanje kakovosti podatkov in ROI, pa je podatkovno skladišče kakovosten vir za podporo odločanju in orodje za povečanje dobička in se mora investicija vanj tudi čim bolje poplačati. Prvi pristop je bolj celovit, drugi pa je usmerjen bolj na dobičkonosnost. Prvi pristop zmanjša tveganja, drugi pa poveča dobiček.

Ne glede na izbrani pristop pa moramo upoštevati, da so podatkovna skladišča z nekakovostnimi podatki le hitra rešitev [9,20], ki jo je treba kmalu dograditi. Projekt razvoja podatkovnih skladišč mora biti izpeljan s spremenljivimi stroški in mora zadovoljiti zahteve uporabnikov ter nuditi dovolj kakovostne podatke.

Pristop k povečanju kakovosti podatkov, zasnovanem na tveganjih, izberemo, ko obstaja:

- zahteva po zelo zanesljivi in brezhibni podpori odločanju,
- želja po zelo kakovostni podpori odločanju,
- z viri dobro podkrepljen projekt.

Pristop k povečanju kakovosti podatkov in ROI izberemo, ko:

- ni zahtevano podatkovno skladišče z zelo kakovostnimi podatki,
- je zahtevan čim večji izplen (povračilo investicije) izdelanega podatkovnega skladišča.

Razvojno metodo DWM uporabimo, ko:

- problem kakovosti ni posebej pereč, ni zahtevana izjemno velika kakovost podatkov,
- kakovost lahko obravnavamo enokovredno z drugimi razvojnimi procesi in izzivi v standardnem zaporedju aktivnosti razvojnega cikla podatkovnega skladišča.

9 Sklep

Izbira metode za merjenje kakovosti podatkov in metode za zagotavljanje in povečanje kakovosti podatkov je pomembna za uspešno izdelavo podatkovnega skladišča.

Za zagotavljanje kakovostnega podatkovnega skladišča je zelo pomemben korak izbire metode merjenja, oziroma ocenitve kakovosti. Pomembno je določiti pravila, ki jih morajo kakovostni podatki zagotavljati. Orodja (kot je Oracle Data Quality Inspector [17]) olajšajo merjenje. Za gradnjo celovitega kakovostnega

podatkovnega skladišča je uporabna obravnavana metoda, ki temelji na tveganjih in nudi metrike za merjenje vseh vidikov kakovosti. Toda realnost je daleč od idelanega sveta in največkrat je pač potrebno meriti in izboljšati tiste lastnosti kakovosti tistih podatkov, ki so potrebne za uspešnost projekta. Kakovost podatkov, ki je manjša kot 90 %, negativno vpliva na poslovanje. Velik korak pa storimo že s tem, če se zavedamo pomena kakovosti podatkov in potrebe po merjenju kakovosti, saj brez merjenja ni možno izvesti izboljšanja in nadzora kakovosti podatkov.

Splošna najboljša metoda za zagotavljanje in povečanje kakovosti podatkov ne obstaja, saj ima vsaka svoje lastnosti. Zato se moramo odločiti za tisto, ki najbolj ustreza dejavnikom v projektu razvoja podatkovnega skladišča. Za brezhibno in zelo zanesljivo podatkovno skladišče uporabimo pristop k povečanju kakovosti podatkov zasnovanem na tveganju. Za projekte razvoja podatkovnih skladišč, ki zahtevajo čim večji ROI, izberemo pristop k povečanju kakovosti podatkov in ROI. Za projekte, kjer lahko kakovost obravnavamo enakovredno z ostalimi razvojnimi izzivi uporabimo metodo DWM.

Ne glede na izbran pristop in razvojna orodja za gradnjo podatkovnih skladišč mora biti projekt izpeljan s sprejemljivimi stroški, zadovoljiti trenutne potrebe uporabnikov, zgraditi dovolj prilagodljivo podatkovno skladišče ter nuditi dovolj kakovostne podatke.

Literatura

[1] Barbusinski, B., Howard, S. Kelley, C. (2002): How would you characterize the difference between Bill Inmon's philosophy on data warehousing and Richard Kimball's?, DM Review.

[2] Current Analysis, (2001): Data Quality Product Assessment.

[3] Dubois., L. (2002): Business Intelligence: The Dirty (and Costly) Little Secret of Bad Data, BI Report.

[4] English, L. P. (1999): Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits, New York: John Wiley & Sons.

[5] English, L. P. (2002): The Essentials of Information Quality Management, DM Review.

[6] Hacknez, D. (2003): Data Warehouse Delivery: Data Quality Fear, DM Review.

[7] Gallas, S. (1999): Kimball Vs. Immon", DM Review.

[8] Groh, T. (2004): Beyond ROI ... Justifying a Business Intelligence Initiative, DM Review.

[9] Global Data Management Survey (2001): PriceWaterhouseCooper, http://pwcglobal.com.

[10] Hufford, D. (1996): Data Warehouse Quality: Special Feature from January 1996.

[11] Immon, W. H. (1996): Building the data warehouse, Wiley, New York.

[12] Kachur, R. J. (2000): The Data Warehouse Management Handbook, Prentice Hall, 2000.

[13] Kimball, R. (1996): The Data Warehouse Toolkit: Practical Techniques for Building Dimenzional Data Warehouses, John Wiley & Sons, New York.

[14] Kimball, R. (1999): The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing and Deploying Data Warehouses, Wiley, New York.

[15] McKnight, W. (2003): Building Business Intelligence: Overal Approach to Data Quality ROI, DM Review.

[16] McKnight, W.: Overall Approach to Data Quality ROI, White Paper, http://www.mcknight-associates.com/.

[17] Oracle Data Quality Inspector: http://www.oracle.com/consulting/offerings/platform/index.html?dqi.html.

[18] Oracle Method (1998): Oracle Data Warehouse Method Handbook, Oracle.

[19] Oracle9i Warehouse Builder 9.2, Integrated Data Quality, http://www.oracle.com.

[20] TDWI Report Series,(2002): Data Quality and the Bottom Line: Achieving Business Success through a Commitment to High Quality Data.

Krista Rizman Žalik ima več kot petnajstletne izkušnje pri analizi podatkov, podatkovnem modeliranju informacijskih sistemov, upravljanju metapodatkov in načrtovanju in razvoju podatkovnih skladišč. Delala je kot projektantka, razvijalka in svetovalka na različnih projektih razvoja informacijskih sistemov in rešitev. V zadnjih petih letih se je posvetila svetovanju, projektiranju in razvoju podatkovnih skladišč.