

Tietovarastointi - tiedonhallinnan painajainen ?

Heikki Kurkinen,
Sybase Finland Oy

Tietovarastointi (Data Warehousing, DW) on yksi tämän hetken kuumimmista aiheista tiedonhallinnan alueella ja sen suosio on helposti ymmärrettävissä. Laajasti ottaen ensimmäistä kertaa tietojenkäsittelyn historiassa painopiste on nyt tiedon hyödyntämisessä tavallisen käyttäjän ja yritysjohdon näkökulmasta sen sijaan, että tiedonhallinta on vain tietojen mahdollisimman tehokkaan tallennuksen kohteena.

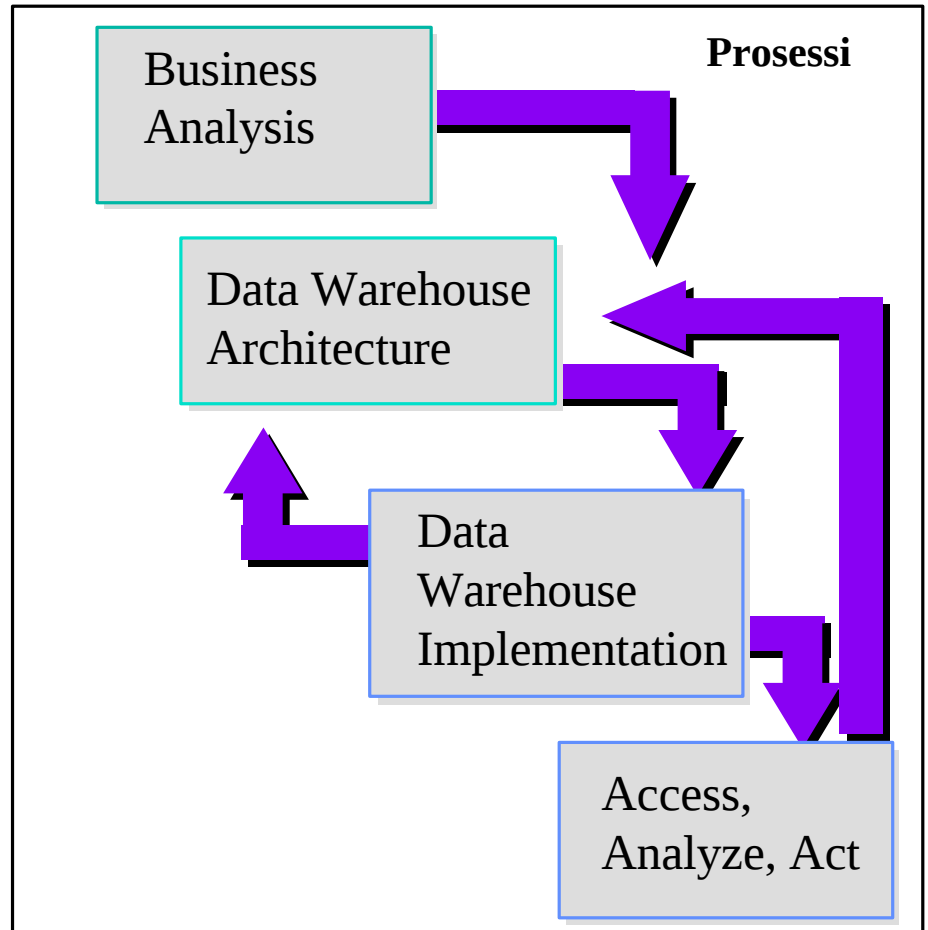
Jo 1980-luvulta lähtien on kehitetty erilaisia johdon informaatio-, EIS jne- järjestelmiä, joiden suosio ei kuitenkaan ole noussut kovinkaan suureksi. Pääsyyt hitaaseen liikkeelleläh- töön ovat olleet järjestelmien otosluonne, tietojen reaaliaikaisuuden puute ja usein hankalat käyttöliittymät tavallisille käyttä- jille.

Nykyisten relaatio- tietokantojen ongelmia

Tietovarastointi on astunut megahypyn kuluneen vuoden aikana mitä tulee tietojen käytet- tävyyteen ja ajantasaisuuteen. Suurimmat syyt ovat varmastikin olleet relaatiotietokantojen leviä- minen käytännön standardirat- kaisuiksi uusissa järjestelmissä sekä kehittyneet kysely- ja analy- sointivälineet. Kuitenkin ongelma- paikkoja on projekteissa selvästi havaittu, joista suinkaan pienin ei ole perinteisen relaatiotietokannan rakenteesta johtuvat vasteajat.

Mikä sitten relaatiokantaratkai-

riippuen yrityksen toimintatavasta



sussa on vaikeata ? Kaikki toimit-
tajat kehuvat kilvan TPC-C bench-
markeilla, joissa nykyään jo
kohtuuhintaisilla laitealustoilla saa-
daan aikaan useiden 10 000
tapahtumien läpimenoja. Ongel-
maksi kuitenkin muodostuu se, että
TPC-C ja muut vastaavat
suorituskykytestit perustuvat mk.
"mixed workload" ympäristöihin,
joissa on sekä päivittäviä että
yksittäisiin riveihin kohdistuvia
kyselyjä. Tietovarastossa tietojen
päivittäminen ei ole jatkuvaa vaan
useimmissa tapauksissa varasto
täytetään eräajoilla 1-2 krt
vuorokaudessa tai viikossa

ja toimialasta. Olennainen kysy-
mys onkin tietojen analysointi,
jossa yksittäisten rivien sijasta
useimmiten haetaan tietojoukkoja,
summia, laskennallisia tuloksia
kuten keskiarvoja, hajontaa, tren-
dejä jne.

Tietokantatekniikan hakuja no-
peuttava indeksointimekanismi ei
ole alunperinkään tarkoitettu tähän.
Tyypillisen asiakasprojektin kulku
on ollut vahvasti pitkään ennakko-
suunnitteluun perustuva, jossa
pyritään arvaamaan etukäteen mitä
käyttäjät tulevat järjestelmästä
kysymään ja mitä valmiita
raportteja tullaan käyttämään.

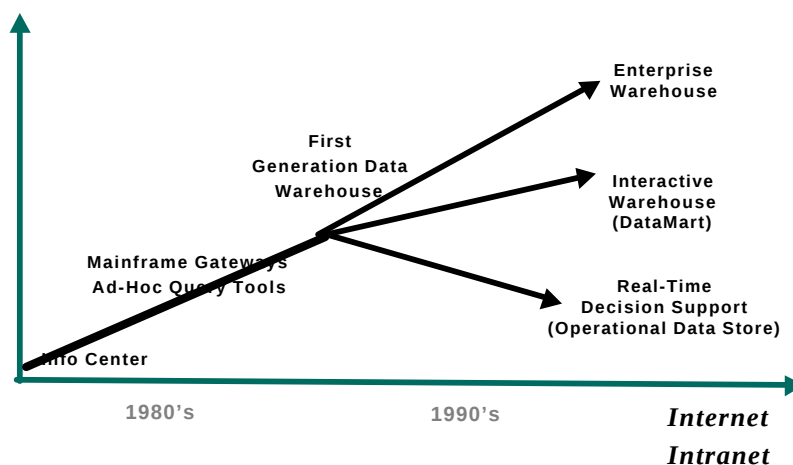
Näiden perusteella järjestelmät ja tiedonhallinta viritetään mahdollisimman tehokkaaksi ja valitettavan usein päädytään sitomaan käyttäjien kädet laskemalla tietojen summatasoja ja ryhmittelyjä valmiiksi. Tällöin menetetään tietovarastoinnin yksi peruskivijaloista, jonka mukaan kaikki yrityksen toimintaan ja ennustamiseen olennaisesti vaikuttava tieto on käsiteltävissä tietovarastossa ilman rajoituksia. Projekteilla on myös paha taipumus venyä ja tulla kalliiksi, yleisimpiä kommentteja on kuulla projektin kestäneen puoli vuotta tai pidempäänkin eikä vielä ole konkreettista, käyttäjille esitettävää toteutusta edes aloitettu. Vastakohtana on sitten projekteja kuten Helsingin Sanomien ilmoitusmarkkinointi, jossa varsinaiseen toteutukseen on tähän mennessä käytetty 15 päivää ja järjestelmä on tuotannossa!

Tietokantojen B-puu indeksointi perustuu avainkenttä ajatteluun, jossa etukäteen osoitetaan MINKÄ kenttien (sarakeiden) mukaan tietokantaan tullaan kohdistamaan hakuja ja MITEN tiedot (taulut) yhdistetään toisiinsa. Perinteinen relaatiokanta on erittäin tehokas yksittäisten rivien käsittelyssä ja ennen kaikkea vastaanottamaan tietoja syövereihinsä, mutta relaatiokanta ei ole optimaalisin ratkaisu silloin, kun tietomassat kasvavat suuriksi ja kun niihin yhdistyy ennalta arvaamattomia hakuja ts. Käyttäjille annetaan vapaammat kädet tehdä tietokantahakuja omien tarpeittensa mukaan.

Systeemyön kannalta kaikki me tietokantojen suunnittelussa mukana olevat tiedämme millainen painajainen tietokannan tehokas indeksointi voi olla. Lisäämällä indeksointia nopeuttaaksemme hakua hidastammme samalla päivityksiä ja ylläpitoa, koska indeksirakenteet on myös pidettävä ajan tasalla muuttuvassa tilanteessa. Samalla tapahtuu nk.

Ratkaisut ovat erilaisia

"One Size does Not Fit All"



10
DW Sales v5 6/95

tietokannan tilanvarausten räjähdys, jota vielä täydentää suorituskykyisistä tehtävä tietojen summaaminen ja ryhmittely valmiiksi. Ylläpito summatason tiedoissa edellyttää myös ohjelmalogiikan toteutusta ja ylläpito entisestään vaikeutuu.

Systeemyöläisen elämää vaikeuttaa vielä CASE-välineiden heikko tuki tietovarastokantojen erikoisrakenteille kuten lumihiutalemalli, tähtimalli ja moniulotteiset rakenteet. Systeemyövälineet ovat ensisijaisesti tarkoitettuja operatiivisten kantojen suunnitteluun ja niiden 'automaattinen' tietorakenteiden (SQL) tuottamiskyky on rajoitettu kolmannen normaalimuodon hyväksyviin tietokantoihin. Tietovarastokannoissa käytön helpottamiseksi tietoa kuitenkin denormalisoidaan ja toistetaan surutta – onhan operatiivisen kannan rakenne aina liian monimutkainen ja hidas-käyttöinen tavalliselle käyttäjälle.

Lisäproblematiikkaa tarjoaa Internet ja sen yhdistäminen tietovarastoon. Erityisesti Intranet-käytössä tietovaraston käyttöliittymänä internet-selain (esim. Netscape) on luonnollinen valinta ja usein suljetut erityisesti

tietovarastokäyttöön rakennetut kannat eivät ole laajalti "Internet-yhteensopivia" tai ainakin edellyttävät ko. toimittajan omien kehitysvälineiden käyttöä merkittävässä määrin.

Vertikaalinen talletusrakenne

Mitä tällä rintamalla sitten on tapahtunut? Perinteiset tietovarastoratkaisut kehittyvät avoimemmiksi ja relaatiokantaratkaisuihin kehitetään kiivaasti uusia innovatiivisia tapoja saada lisää suorituskykyä, mutta perusrakenteen aiheuttamia pullonkauloja tällä ei poisteta. Luonnollisesti lisäämällä laitekapasiteettia ongelmaa saadaan siirrettyä eteenpäin, mutta kustannukset alkavat sitten hipoa tähtitieteellisiä lukuja ja perusteluja toimivalle tietovarastolle on haettava lisää.

Tietovarastointiin on myös kehitetty relaatiokantarajapintoja käytäviä tiedonhallintaratkaisuja, joista esimerkkinä mainittakoon Sybase IQ. Kääntämällä perinteinen relaatiorakenne päällelleen voidaan toteuttaa tietovarastoja, joissa vasteajat putoavat jopa tunneista minuutteihin ja minuuteista sekunteihin. Lisäksi ennalta-

SELECT COUNT(*) FROM CARDHOLDER WHERE CARD="VISA" AND STATE IN ("CA", "MA")

Perinteinen relaatiokanta

▲
10 Milj
riviä
▼

Card	State	Name
VISA	NY	
MC	CA	
AMEX	CT	
VISA	MA	
-	-	-

← 200 tavua / rivi →

$\frac{10\,000\,000 \times 200 \text{ tavua}}{16 \text{ Ktavua}}$

125,000 I/O:ta

Sybase IQ

▲
10 Milj
bittia
▼

VISA	State=CA	State=MA
1	0	0
0	1	0
0	0	0
1	0	1
-	-	-

$\frac{3 \times (10\,000\,000 / 8) \text{ tavua}}{16 \text{ Ktavua}}$

235 I/O:ta

arvaamattomat kyselytarpeet voidaan toteuttaa tehokkaasti; nyrkkisääntö on jopa se, että mitä monimutkaisempi kysely on sitä nopeampi vasteaika verrattuna perinteiseen ratkaisuun.

Sisäinen rakenne on vertikaalinen eli rivien sijasta tietokannassa onkin toisistaan riippumattomia olioita (sarakkeita), jotka kaikki ovat indeksoituja ja lisäksi pakattuja levy i/o:n minimoimiseksi. Projekteissa on usein päästy jopa usean sadan tietokantataulun ratkaisuihin yhden tai korkeintaan muutaman taulun rakenteisiin. Rakenne on myös siitä miellyttävä, että se on helppo toteuttaa ja käyttäjille ymmärrettävä.

Jo tietojenkäsittelyn kannalta ammoisina aikoina kehitetty bittikarttaindeksointi soveltuu erinomaisesti vertikaalisen talletusrakenteeseen, koska sarakkeet elävät tietokannan sisällä omaa elämäänsä – rivikäsitettä ei perinteisessä mielessä ole. Bittikarttojen rajoituksena on usein pidetty sen kykyä esittää tehokkaasti sellaisia sarakkeita, joissa arvojen vaihtelu on enemmän kuin kymmenen ts. arvoalueet kuten sukupuoli (M,N) on bittikartalle helppoa mutta jos arvoalue on useita satoja (Suomen kunnat, postinumeroalueet, teleoperaattorialueet jne. bittikarttaa ei ole pidetty soveliaana vaihtoehtona.

Nykyteknologialla tämäkin rajoitus on poistunut ja esimerkiksi Sybase IQ:n bittikarttamekanismi on jopa saanut USAssa patentin innovatiivisuudestaan.

Kun tähän yhdistetään CASE-työväline, joka pystyy lähes mistä tahansa tietokannasta imuroimaan suoraan tietokantarakenteet ja tuottamaan hallitusti tietovarastotietokannan rakenne, on systeemyökin huomattavasti helpompaa kuin integroimattomissa tai suljetuissa ratkaisuissa. Asiakasprojektien yksi merkillepantava piirre on myös mahdollisuus käyttää useampaa eri kyselyvälinettä tietovarastoon: taulukkolaskenta yksinkertaisimpiin tarpeisiin, valmiit raportit 4GL-kehittimillä ja sitten todelliseen tehokäyttöön monipuoliset analysointityökalut kuten esimerkiksi BusinessObjects.

Mitä olemme oppineet projekteissa ja asiakkailta? Tietovarastointi on ehdottomasti varteenotettava vaihtoehto tiedon hyödyntämiseen ja operatiivisten järjestelmien kuormituksen pienentämiseen. Teknologia-avainnat tänä päivänä ovat joko erilaisiin suljettuihin kuutiomalleihin, perinteisiin relaatiokantamalleihin tai puhtaasti tietovarastointiin tarkoitettujen tietovarastojen kesken tehtäviä päätöksiä. Kuitenkin periaate "one-size-fits-all" on erityisesti tieto-

varaston kohdalla osoittautunut heikoksi ja huonosti kantavaksi ratkaisuksi, koska tietovarastointi on aiemmasta täysin poikkeavia vaatimuksia esittävä kokonaisuus. Useasta komponentista koottava ratkaisu, jossa valitaan kuhunkin tasolle se parhaiten sopiva osa, on kuitenkin jatkokehityksen ja ennen kaikkea sen tärkeimmän eli käyttäjien hyväksynnän saamiseksi paras mahdollinen vaihtoehto – käyttäjät ja tietovaraston tapauksessa korostetusti yrityksen ylin johtohan se kuitenkin maksaa projektin ja haluaa myös rahoilleen selkeän, mitattavan vastineen: hyödyn liiketoiminnan kehittämiseksi ja kilpailuedun.

Keskustelen mielelläni tietovarastojen toteutuksesta ja jaan kokemuksiani tuottavista ratkaisuista:

*Heikki Kurkinen,
Sybase Finland Oy,
Heikki.Kurkinen@sybase.com
tai puh. 09-7250 2288.*