

Tietojen summaus ja jalostus

Ari Hovi,
Ari Hovi Oy

Esittelen Datawarehouse-kantojen summaamisen ja tietojen jalostuksen etuja ja ongelmia. Artikkelini on otos kirjastani "Data Warehousing - tietovarastotekniikka", Suomen atk-kustannus.

Tässä kappaleessa esittelen tietojen summaamisen ja johdettujen tietojen etuja ja ongelmia.

Summataulut

Varsinainen "taikasauva" tietovarastotekniikassa on **summaaminen**. Se täydentää hyvin tähtimallia (star schema). Summauksella tietovaraston tietoja jalostetaan korkeammalle jalostusasteelle ja selätetään samalla monia tietovarastojen ongelmia.

Summatauluista on eniten hyötyä sellaisten summien kohdalla, joita paljon kysellään. Jos jotakin summaraportti ajetaan kerran kuussa, ei ehkä kannata rakentaa summataulua sitä varten. Jos taas jotakin summausta tarvitaan joka päivä kyselyissä ja raporteissa, ei kannata laskea samoja summia aina uudestaan ja uudestaan vaan lasketaan ne valmiiksi.

Perustauluilla tarkoitan tietovaraston yksityiskohdiltaan alimman tason tauluja, jotka ovat yleensä isoja. Isoista perustauluista muodostetut **summataulut** ovat huomattavasti pienempiä, yhdistelemisen ja denormalisoinnin (tietoja toistetaan useaan

tauluun) ansioista liitostarve pienenee ja lisäksi summa-
tauluihin laitetaan johdettuja tietoja eli kaavoilla perustiedoista jalostettuja tietoja.

Erilaiset asiakastuotot tai katteet ovat esimerkkejä **johdetuista tiedoista**.

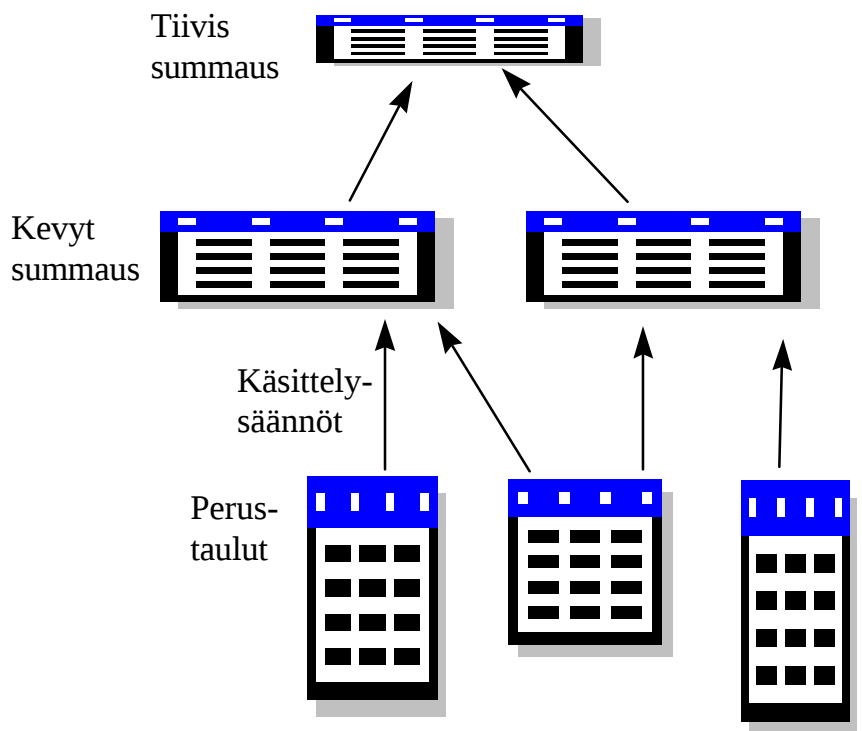
On parempi sijoittaa johdetun tiedon laskenta summauksen yhteyteen eikä erilaisiin raporttiohjelmiin.

Johdettu tieto ja siinä tarvittava kaava ikäänkuin julkaistaan. Näin perustietojen lisäksi myös johtamiskaavat ja yhdenmukaiset johdetut tulevat kaikkien käyttöön (tietysti valtuuksien

puutteissa).

Usein tullaan toimeen kevyellä summauksella, sillä nimes-
tään huolimatta tässä summauk-
sessa tiedot tiivistyvät usein kym-
meneen, jopa yhteen prosenttiin
perutaulun koosta. Tarvittaessa
voidaan kuitenkin tehdä vielä
tiiviimpi summaus ja jalostus.
Tällaisia tietoja voisivat olla
johdolle tiivistetyt liiketoiminnan
tärkeät tunnusluvut, joita ei enää
ole kovin paljon, mutta jotka
sisältävät paljon informaatiota.

Summaaminen on yksi tärkeimmistä menetelmistä tietovarastotekniikassa.



Eri tasoisia summauksia

Summaamisella saavutetaan monia etua :

1. **Kyseltävä massa pienenee**, koska summataulussa on huomattavasti vähemmän rivejä kuin perustauluissa. CPU ja I/O-kuorma pienenevät, jolloin kyselyt nopeutuvat, yleensä huomattavasti.
2. Summataulut luodaan usein **monesta taulusta** liitoksella, siis denormalisoiden. Kun summatauluun on valmiiksi koottu monen taulun tietoja, yksinkertaistuvat kyselyt huomattavasti ja myös nopeutuvat.
3. Yleisesti tarvittavat, perustiedoista **johdetut tiedot** lasketaan summaamisen yhteydessä ja ne sijoitetaan valmiiksi laskettuina summatauluihin. Kun johtamiskaavasta on sovittu, on kaikilla nyt samat keskeiset tunnusluvut käytävissä. Käyttäjien ei tarvitse aina uudestaan laskea samoja lukuja rasittaen suorituskykyä.
4. **Monimutkaiset summaamiset** tehdään keskitetysti ja oikein (käyttäjä ei pääse tekemään itse outoja yhdistelmiä).
5. Päästään hyödyntämään helppokäyttöisiä ja voimakkaita raporttigeneraattoreita ja kyselyvälineitä; parhaimmillaan haut tulevat yhdestä taulusta (kts. kohta 2) eikä raporttiohjelmassa enää tarvita juurikaan laskentakaavoja tai käsittelylogiikkaa (kts. kohta 3), lähinnä vain hakuehtoja.

Käytännön esimerkki summaamisesta

Päivätasolla oleva kirjanpito-aineisto oli raportoinnin ja kyselyjen kohteena. Kannan koko oli n. 500 M ja se sijaitsi minikoneella. Käyttäjät tekivät kuun vaihteen jälkeen parametroitavalla raporttigeneraattorilla kyselyjä aineistoon. Kun kyselyjä kasaantui koneelle, alkoivat vastausajat kasvaa, jopa tunneiksi.

Asiaa tutkittaessa kävi ilmi, että lähes 90 % käyttäjien raporteista kohdistui kuukausitasolle, vaikka tarjolla oli myös päivätasoa. Siis raporttijaissa tehtiin paljon kuukausitason summaamista.

Ongelman ratkaisuksi luotiin

aineistosta kuukausitason summataulu, jonka kooksi tuli vain n. 50 M. Summataulun tekeminen tapahtui yhdellä SQL-lauseella (INSERT - SELECT) ja sen luonti kesti 30 min. Summataulu ajettaisiin kerran kuussa.

Summatauluun kohdistuvat raportit kestivät nyt vain pari minuuttia eivätkä ne enää kasaantuneet tukkimaan konetta. Summataulun hinta oli siis aika vähäinen : 50 M levytilaa sekä 30 min ajo kerran kuussa. Edut sen sijaan olivat huomattavat.

Summauksen perusvaihtoehdot

Tarkastelen tässä muutamaa erilaista vaihtoehtoa laatia summataulut.

Erilliset taulut vai osa perustauluja

Tavanomaisin ratkaisu on laatia summataulut erillisiin, omiin tauluihinsa. Tauluja syntyy nyt lisää, mutta ne ovat suhteellisen pieniä helpottaen täten niiden hallintaa.

Toinen vaihtoehto joissakin tapauksissa on tallettaa summätietueet samaan tauluun kuin perustiedotkin. Tällöin mukaan on talletettava tasoa kuvaa tieto, jotta kyselyn voi kohdistaa vain summariveihin tai vain perustietoriveihin. Etuna on se, että taulujen määrä pysyy pienempänä. Haittana on se, että taulujen koko entisestään kasvaa, jolloin summataulujen pienuusetua ei saavuteta.

tuotenro	liike_nro	mk	edkk_mk	v_al_mk	ed_v_al_mk
32	14	24355	24320	132333	132300
32	12	34253	34210	144201	144248
32	8	43263	44100	169533	145888
...	...	...	...	...	...

Taulu 1: Esimerkki kuukausitason yhteenvetosummataulusta.

Yhteenvetosummataulu

Yhteenvetosummataululla

tarkoitaa nopeata, vähän tilaa vievää pikaratkaisua, joka hyvinkin saattaa täyttää käyttäjien akuuteimmat tietotarpeet. Ideana on se, että tällaiseen summatauluun viedään mahdollisimman paljon käyttäjiä kiinnostavaa summatietoa, johdettuja tietoja ja myös erinäisiä historia-summatietoja, mutta tämä historia viedään vain koottuna summatasolla. Varsinaista historiaa siihen tapaan kuin "oikeaan" tietovarastoon talletetaan ei kuitenkaan talleteta.

Taulussa 1 on tuotteittain ja liikkeittäin viime kuukauden myynti sekä valmiiksi laskettuna edellisen kuukauden myynti

(edkk_mk), vuoden alusta myynti (v_al_mk) sekä viime vuoden alusta vastaavaan kuukauteen myynti (ed_v_al_mk). Tyypillinen kenttä olisi vielä edellisen vuoden saman kuukauden myynti.

Tällainen taulu on hyvin nopea ja kätevä kysellä. Siihen voidaan koota käyttäjien kaikkein yleisimmät tarpeet, kuten juuri vertailut edelliseen kauteen, viime vuoden samaan kauteen ja vuoden alusta -yhteensä luvut. Tauluun ehkä myös toistetaan tuotteiden ja liikkeiden nimet mukaan, mikä kyllä vie tilaa mutta helpottaa raporttien tekoa.

Taulu pysyy melko pienenä, sillä se ei kasva ajan kuluessa, kuten varsinaiset tietovaraston taulut, joihin aina lisätään uuden

kauden tiedot. Hintana tästä kompaktista koosta on se, että taulusta ei saa toissakuukauden myyntiä eikä muita vastaavia spontaaneja kyselyjä. Taulun rakentaminen on myös suhteellisen monimutkaista ja esimerkiksi edellisen vuoden summien mukaan saaminen edellyttää, että historiatiedot ovat kuitenkin jossakin tallella.

Summaamisen ongelmia

Mitalilla on kaksi puolta. Edellä mainittujen sangen merkittävien etujen hintana on pohdiskeltavanamme muutamia seikkoja.

Summaamisen ongelma-alueita

1. Millä dimensioilla summataan.

- summatauluvaihtoehtoja saattaa tulla paljon

2. Summaus vie tilaa (jos perustaulutkin jätetään)

- tosin summataulut ovat kooltaan pienempiä

3. Summaaminen vie aikaa

- riittävätkö tunnit tehdä latausten jälkeen vielä summaukset

4. Summaus lisää kannan monimutkaisuutta

- tauluja tulee lisää ja summausohjelmat on laadittava

5. Summatuista taulusta menetetään yksityiskohtia

- kuukausisummataulusta ei tehdä päivätason analyysijä

6. Tietojen eheyden kanssa pitää olla tarkkana

- kun tietoja toistetaan (summaaminenkin on toistoa), on aina enemmän vaaraa, että tiedot menevät epätahtiin.