Automaatio 1999, 14.-16.9.1999, Helsinki.

OLAP-tekniikoiden käyttömahdollisuudet teollisuusprosessien analysoinnissa

Jukka Kiviniemi, Antoni Wolski
VTT Tietotekniikka, PL 1201, 02044 VTT, Finland
Puh. (09) 456 5936, 456 6012, Fax. (09) 456 6027, sähköposti: {jukka.kiviniemi,antoni.wolski}@vtt.fi

AVAINSANAT analysointi, tietovarastointi, OLAP, aikasarjat

1. JOHDANTO

Tietovarastointia (data warehousing) [Hov97] on viime aikoina alettu hyödyntää yritysten taloudellisissa analyyseissä. Yritysten operatiiviset tietokannat eivät usein täytä käyttäjien raportointitarpeita riittävän kattavasti. Raporttien teko vaatii usein sovellusohjelmointia ja vaikka tehokkaita raporttigeneraattoreita olisikin käytettävissä, on tietokannan sisäinen rakenne usein sellainen, että tehokas ja nopea raportointi ei ole mahdollista. Operatiivisten tietokantojen päätarkoituksena on suorassa käytössä syötettyjen transaktioiden (tapahtumien) tehokas käsittely. Tietovarastoinnin ideana on kerätä yrityksen tietokannoissa oleva tieto yhteen tietovarastoon niin, että joustava raportointi on mahdollista. Tietovarastointia ja siihen liittyvää raportointia kutsutaan OLAP-tekniikaksi (on-line analytical processing) [CCS93]. OLAP ei ole mikään yksittäinen ohjelmisto tai menetelmä, vaan se kattaa suuren joukon menetelmiä tiedon talletuksesta raportointiin ja visualisointiin.

Teollisuusprosessien analysoinnissa on tarvetta sekä prosessin ajonaikaiseen että myöhemmin tapahtuvaan analysointiin. Perinteinen OLAP-teknologia soveltuu kohtalaisen hyvin myöhemmin tapahtuvaan analysointiin mutta ajonaikaisten analyysien suorittaminen on ongelmallista (OLAP-lyhenteen sana "on-line" viittaa tietovarastossa olevien tietojen käyttötilanteeseen eikä ajonaikaisten tietojen saatavuuteen). Tietovarastot päivitetään eräajoina silloin kun järjestelmän muu käyttö on vähäistä, esimerkiksi öisin. Toinen ongelma on se, että tietovaraston koon kasvun estämiseksi tietojen aikatarkkuutta usein pienennetään. Esimerkiksi mittausaikasarjoista ei talleteta todellisia lähtöarvoja vaan tuntikohtaiset keskiarvot. Kolmantena ongelmana on se, että aikaan kohdistuvat kyselyt ovat usein tehottomia ja perinteiset OLAP-menetelmät eivät tarjoa tähän ongelmaan selvää ratkaisua.

Tässä artikkelissa esitellään OLAP-tekniikan perusperiaatteet ja tekniikoiden soveltamismahdollisuuksia teollisuusprosessien analysoinnissa. Luvussa 2 käydään läpi OLAP-tekniikan perusteet sekä esitellään eräitä kaupallisten OLAP-tuotteiden toimittajia. Luvussa 3 käydään läpi teollisuusprosessien analysointitarpeita. Luvussa 4 pohditaan, millaisia laajennuksia nykyinen OLAP-tekniikka vaatii, jotta sitä voitaisiin soveltaa teollisuusprosessien analysoinnissa.

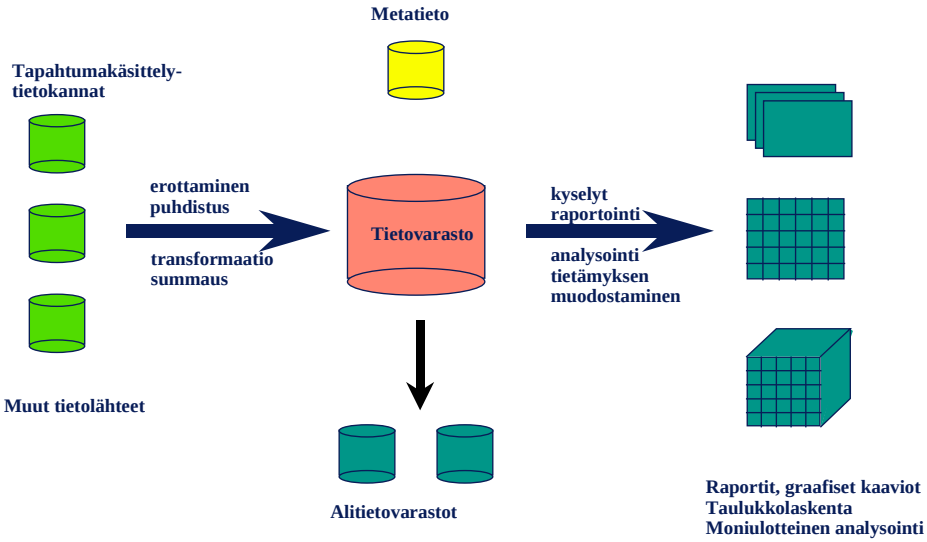
2. OLAP-TEKNIIKAN PERUSTEET

OLAP on prosessi, joka sisältää tietovarastointiin ja tiedon analysointiin liittyviä vaiheita. Yleensä tietovarasto on yritysten operatiivisista järjestelmistä irrallaan oleva tiedon talletuspaikka, jota käytetään ainoastaan analysointitarkoituksiin. Suurin osa tietoa siirretään tietovarastoon operatiivisista tietojärjestelmistä (esim. yritysten tapahtumakäsittelytietokannat) yhteiseen. Tietovarastoon voidaan siirtää tietoa myös muista tietolähteistä, kuten tietoja säätilasta, pörssikursseista jne. Useiden erityyppisten tietolähteiden integrointi samaan tietovarastoon vaatii sen, että tietoa esikäsitellään ennen tiedon siirtoa. Tyypillisiä esikäsittelyvaiheita ovat relevantin tiedon erottaminen (extraction), ylimääräisen ja toisteellisen tiedon siivous (cleansing), esitysmuotojen muunnokset (transformation) sekä tietojen koostaminen (summarization, aggregation).

Hyvä tietovarasto tarjoaa samassa paikassa kaikki yrityksen toiminnan kannalta relevantit tiedot analysointia varten. Joskus on kuitenkin tarvetta esimerkiksi tietosuojasyistä johtuen erottaa osa informaatiosta omaan alitietovarastoon (data mart). Tietovarastoon liittyy oleellisena osana metatieto (metadata), jossa kuvataan tietovaraston rakenne käyttäjiä varten.

Tietovarastossa olevia tietoja voidaan analysoida erilaisilla menetelmillä. Tietoon voidaan kohdistaa helposti joustavia kyselyitä ja siitä voidaan ajaa monimutkaisia raportointeja. Samanlaisten vaa-

timusten täyttäminen operatiivisesta tietokannasta käsin olisi erittäin hankala ja tehotonta mm. siksi, että tiedot siellä eivät ole koostettuja. Tietoja voidaan hakea taulukkolaskentaohjelmiin ja niitä voidaan analysoida erilaisilla moniulotteisilla analysointityökaluilla. Lisäksi tietovarastossa olevan tiedon avulla voidaan etsiä uutta tietämystä yrityksen toiminnasta (data mining). Kuvassa 1 esitetään OLAP-järjestelmän yleisarkkitehtuuri.

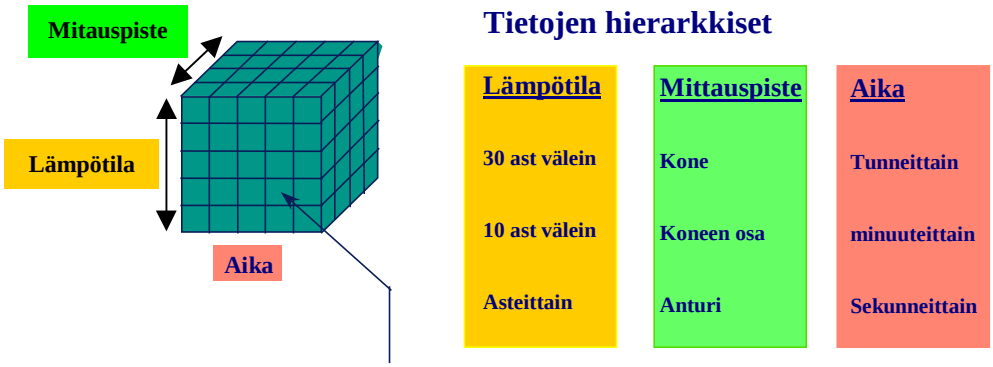


Kuva 1. OLAP yleisarkkitehtuuri.

Tietovarastossa oleva tieto eroaa operationaalisessa tietokannassa olevasta tiedosta pääasiassa siinä, että tietovaraston tieto on valmiiksi koostettua. Koostaminen voidaan tehdä eri menetelmillä. Tyypillisiä menetelmiä ovat yhteenlasku, keskiarvoistaminen sekä havaintojen määrän laskeminen. Keräyttäminen voidaan tehdä useilla eri tasoilla, esimerkiksi jos tietovarastossa säilytetään tietoa prosessimittauksista, voidaan tiedot koostaa esimerkiksi mittauspisteittäin, koneittain, koneen osittain sekä aikaskaalassa tunneittain tai minuuteittain. OLAP-tekniikka tarjoaa tehokkaan tavan käsitellä eri summaustasoilla olevia tietoja, koska tulokset ovat valmiiksi laskettuja.

Tietovaraston sisältöä mallinnetaan usein datakuutiona (data cube) [GBLB96]. Datakuutio on moniulotteinen analyysimalli, joka koostuu dimensioista (dimension) ja faktoista (fact). Dimensiot muodostavat moniulotteisen ristiintaulukoinnin. Fakta tarkoittaa dimensioiden leikkauspisteessä olevaa tietoa, joka lasketaan halutulla koostefunktiolla (aggregate function). Datakuutioon voidaan kohdistaa useita erilaisia operaatioita. Kuutiota voidaan kääntää (rotate), jolloin tiedon tarkastelukulma vaihtuu. Lisäksi kuutiosta voidaan valita haluttuja osia (slicing, dicing) tarkastelun kohteeksi. Kuutio voi sisältää jokaisesta dimensiosta useita hierarkkisia tasoja, joita voidaan käsitellä porautumalla (drill down) tai yleistämällä (roll up). Porautumisessa tarkastellaan samaa tietoa suuremmalla tarkkuudella. Esimerkiksi jos tehdään analyysi tunneittain, voidaan porautumalla saada vastaava analyysi minuuteittain. Porautumisen maksimitarkkuutta rajoittaa se, millä tarkkuudella tiedot on talletettu tietovarastoon.

Kuvassa 2 on esimerkki kolmiulotteisesta datakuutioista. Kuutioon on talletettu tieto siitä, kuinka usein eri mittauspisteessä on esiintynyt tiettyyn arvovälin kuuluva lämpötila-arvo eri aikoina. Kuution dimensiot ovat siis lämpötila, mittauspiste ja kellonaika. Lämpötilasta on tarjolla 3 eri skaalausta, 30 asteen välein, 10 asteen välein ja asteittain. Vastaavasti myös muista dimensioista on tarjolla eri tarkkuudella olevia tietoja. Kuution tieto on leikkauspisteessä esiintyneiden havaintojen määrä. Kuvan alaosan taulukossa esitetään, millaiselta esimerkissä mainitun datakuution yksi näkymä voisi käytännössä näyttää. Siinä mittauspistedimensio on koostettu (tässä tapauksessa dimensiossa olevat arvot on laskettu yhteen, jolloin SUM()-koostefunktiota on käytetty. Taulukot sarakkeet ovat aikavälejä ja rivit lämpötila-alueita.



Fakta: kuinka usein lämpötila-arvo on esiintynyt kussakin pisteessä

	8:10	8:20	8:30	8:40
100	13	11	11	8
110	25	30	25	28
120	23	0	14	17
130	14	13	12	18
140	11	14	20	22

Kuva 2. Datakuutio prosessianalyysissä.

Datakuution tiedot voidaan relaatiotietokannassa esittää nk. faktataulun avulla. Kuvan 2 faktataulu on seuraavan kaavan mukainen:

faktataulu:(mitt_piste_nro, lampotila, aika_10min)

Nykyinen relaatiotietokantojen SQL-standardi [ISO92] ei mahdollista koko datakuution kuvaamista yhdellä kyselylauseella vaan jokainen kuution näkymä täytyy kuvata erikseen. Esimerkiksi kuvan 2 taulukon esittämä tahko voidaan kuvata SQL-lauseella

```
select lampotila, aika_10min, count(*)
from faktataulu
group by lampotila, aika_10min;
```

Tosin tällä komennolla tuotettu tulostaulu ei ole kuvan 2 taulukon näköinen, vaan koostearvot ovat siinä allekkain:

```
lampotila   aika_10min   count(*)
-------------------------------
100         8:10         13
100         8:20         11
100         8:30         11
(...)
110         8:10         25
110         8:20         30
(...)
```

Jokainen yksittäinen näkymä kuvataan varioimalla group by-lauseen attribuutteja. Tällöin koko kuution kuvaamiseen standardilla SQL:llä vaaditaan 2^n SQL-lausetta, missä n on datakuution dimensioiden määrä. Käytännön datakuutioissa, esimerkiksi kun dimensioita on 20, laskennan määrä kasvaa liian suureksi nykyaikaisella tietokonekapasiteetilla toteutettavaksi. Tämän vuoksi kuution laskentaan on kehitetty useita erilaisia optimointimenetelmiä, jotka osaavat hyödyntää välituloksia kuution seuraavien osien laskennassa [AAD96].

Lukuisat järjestelmätoimittajat tarjoavat OLAP:iin ja tietovarastointiin soveltuvia ohjelmistoja. Seuraavassa esitetään tästä lukuisasta ohjelmatarjonnasta eräitä esimerkkejä. Tietovarastoinnin

uranuurtajiin kuuluu Suomessa vähemmän tunnettu nykyään Informixin omistama Red Brick[1]. Tyypillisesti jokaisella suurimmalla tietokantatoimittajalla on tietokantaohjelmistossaan tarjolla tietovarastointimahdollisuus esimerkkinä Oracle Warehouse[2], Sybase IQ[3] ja Microsoft SQL Server 7.0[4]. Esimerkki pelkästään moniulotteiseen analysointiin tarkoitetuista ohjelmistoista on kotimainen Voyant[5]. Eräät tilasto-ohjelmistojen valmistajat ovat tuoneet ohjelmistoonsa mukaan tietovarastointi- ja analysointimahdollisuuksia, esimerkkinä SAS[6].

3. TEOLLISUUSPROSESSIN OLAP-ANALYSIN ESIMERKKI

Seuraavassa esimerkki siitä, miten OLAP-tyyppisiä analysointiominaisuuksia voitaisiin hyödyntää prosessinhallintajärjestelmissä.

Esimerkki: Paperikoneiden sähkökäyttöjen koostearvojen tarkastelu.

Paperitehtaalla on useita paperikoneita. Koneiden sähkökäyttöjen (moottoreiden) tilaa seurataan tosiaikaisesti. Moottoreiden lämpötila ja kuormitusaste (hetkellinen teho / nominaaliteho) mitataan ja keskiarvoistetaan aikavälein 1 min, 5 min, 10 min, 30 min ja 1 h jokaisen moottorin kohdalla. Esimerkissa vaatimuksena on mahdollisuus tarkastella sähkömoottorien tilaa eri moottorikokoelmissa kokoelmakohtaisten keskiarvojen avulla. Esimerkiksi halutaan:

A. Vertailla keskenään moottorien kuormitusastetta paperikoneen eri osien välillä, (kuten märkäpää, kuivaaja ja päällystäjä) ja vertailla tilannetta eri koneiden välillä; tässä tapauksessa kuormitusastearvot ovat keskiarvoistettuja jokaista paperikoneen osaa kohti ja laskennassa käytetään esimerkiksi 10 min arvoja.

B. Nähdä edellä esitetyt arvot esim. 1 min tai 1 h aikaresoluutiolla (porautuminen ja yleistäminen aikadimensiossa)

C. Nähdä edellä esitetyt arvot joko pienempien moottoriryhmien kohdalla tai koko koneen tunnuslukuna (porautuminen ja yleistäminen fyysisen sijainnin dimensiossa).

D. Vertailla lämpötilan keskiarvot eri moottorityyppien tai teholuokkien välillä.

Jotta edellä kuvatut raportointitavat tulisivat mahdollisiksi, täytyy jokaisen moottorin lämpötilan ja kuormituksen keskiarvon kohdalla olla analyysiavaruuden (datakuution) koordinaattitieto. Esimerkin analyysidimensiot ovat: fyysinen sijainti (eri tarkkuudella yksittäisestä moottorista koko tehtaaseen), aika (eri tarkkuudella), moottorin tyyppi ja teholuokka. Näitä koordinaattiarvoilla varustettuja lähtötietoja kutsutaan faktatiedoiksi ja sitä vastaavaa rakennetta faktatietokuutioksi. Tässä esimerkissä syntyy kaksi faktatietokuutiota: kuormituskuutio ja lämpötilakuutio

Faktatietojen pohjalta lasketaan koostamalla yhä tiivimpiä datakuutioita, joiden dimensioiden määrä vähenee sitä mukaa kun jokin dimensio poistuu koostamisen tuloksena. Esimerkiksi, kun lasketaan datakuutio, jossa keskiarvot on laskettu teholuokasta riippumatta (kuten tapauksessa A.), teholuokkadimensio poistuu. Koostamista voidaan jatkaa, kunnes saadaan jokaisen faktatietokuution kohdalla yksi arvo. Esimerkissä nämä arvot olisivat: koko tehtaan moottoreiden kaikkien aikojen keskiarvokuormitusaste ja vastaava keskiarvolämpötila. Todellisuudessa esimerkin kaltaisessa prosessivalvontasovelluksessa aikadimensio ei poistu, vaan kaikki koosteet halutaan nähdä aikasarjoina. Tässä tapauksessa läpikotaisin koostettu tieto (koko tehtaan keskiarvokuormitusaste ja keskiarvolämpötila) esitettäisiin eri aikaresoluution mukaisina aikasarjoina.

Suurin ero tämän esimerkin mukaisella OLAP-analyysillä ja perinteisellä OLAP-analyysillä on, että perinteisessa tapauksessa tietovarasto ei muutu käytön aikana. Esimerkin tapauksessa toimiva

[1] http://www.redbrick.com
[2] http://www.oracle.com/datawarehouse
[3] http://www.sybase.com/bid
[4] http://www.microsoft.com/SQL
[5] http://www.brossco.fi
[6] http://www.sas.com/software/data_warehouse

prosessi tuottaa jatkuvasti uusia mittaustietoja ja niiden vaikutuksen pitäisi näkyä välittömästi analyysituloksissa. Jokainen faktatiedon (esim. 1 min keskiarvo) muuttaminen johtaa eritasoisten datakuutioiden uudelleen laskemiseen, koska muutos propagoituu läpi kuutiorakenteen. Jotta tästä vaativasta laskentatarpeesta voitaisiin suoriutua, täytyy ottaa käyttöön uudenlaisia optimointimenetelmiä, jotka vähentävät välittömän koostearvojen uudelleenlaskennan tarvetta.

4. OLAP-TEKNIIKAN SOVELTAMINEN ANALYYSISSÄ

Teollisuusanalyysi asettaa OLAP-tekniikoille lukuisia erikoisvaatimuksia. Perinteiset OLAP-tekniikat on suunniteltu yritysten kaupallisia analyysitarpeita silmälläpitäen. Teollisuuden analyysitarpeiden täyttämiseksi OLAP-tekniikkaa tulee kehittää edelleen. Seuraavassa analysoidaan tärkeimpiä kehityskohteita. Oletamme, että prosessista laaditaan jonkinlainen analyysimalli (analysis model) joka toteutetaan datakuutiona. Analyysimalli voi sisältää useita seurattavia prosessimuuttujia joissa jokaisessa voi olla useita tarkkuustasoja. Lisäksi analyysimalli voi sisältää useita erilaisia koostefunktioita.

Perinteisesti OLAP-tekniikka ei tarjoa mahdollisuuksia reaaliaikaiseen analyysiin. Datakuution päivitys on raskas laskennallinen toimenpide, minkä johdosta päivitys suoritetaan silloin, kun järjestelmä ei ole muuten käytössä. Usein prosessia on tarvetta analysoida reaaliaikaisesti, esimerkiksi kun näytetään operaattoreille monimutkaisia trendinäyttöjä. Lisäksi prosessista mitattavan tiedon määrä ja saapumisnopeus on usein suuri. Jatkuvatoimisessa prosessissa ei ole mahdollista määritellä sellaista ajanjaksoa, jossa datakuution päivitys eräajomuotoisena olisi mahdollista. Reaaliaikainen päivitys vaatii uudenlaisia lähestymistapoja, joita on viime aikoina kehitetty VTT Tietotekniikassa [KWPA99]. Lähestymistavassa lähdetään siitä, että teollisuusanalyysin mittaustiedot ovat luonteeltaan epätarkkoja minkä johdosta myös analyysimallille sallitaan tietty epätarkkuus. Epätarkkuuden salliminen pienentää laskentatarvetta merkittävästi ja mahdollistaa datakuution reaaliaikaisen päivityksen kun prosessista saapuu jatkuvasti uutta tietoa.

Ajan käsittely analyysimallissa voi olla ongelmallista. Aikaan perustuvat moniulotteiset kyselyt ovat perinteisillä OLAP-tekniikoilla tehottomia. Teollisuusanalyysissä on usein tarvetta käsitellä aikaa useissa eri skaaloissa. Lisäksi usein on tarvetta käsitellä samaa mittaustietoa useilla eri aikasemantiikoilla. Tyypillisesti tärkeitä aikakäsitteitä ovat aika, jolloin mitattava ilmiö on todellisuudessa tapahtunut sekä aika, jolloin mittaustulos on kirjattu tietokantaan. Analyysimalleja voidaan käyttää myös lyhytaikaisten ennusteiden laatimisessa. Teollisuuden OLAP-tekniikan tulee tarjota menetelmiä ennusteiden laatimiseksi sekä ennustettujen tulosten ja toteutumatietojen vertailuun.

Teollisuuden mittaustieto on harvoin täysin virheetöntä. Yksittäiset arvot voivat puuttua, olla virheellisiä tai arvot ovat käytettävissä analyysin kannalta liian myöhään. Puuttuvien arvojen käsittelymenetelmät voidaan jakaa seuraavaan neljään pääluokkaan

1. ei välitetä puuttuvista arvoista lainkaan
2. interpoloidaan puuttuvat arvot (kyselyiden vastauksia muodostettaessa)
3. korvataan aikasarjojen puuttuvat arvot muiden aikasarjojen avulla
4. annetaan puuttuvien arvojen vaikuttaa analyysin tuloksen tarkkuuteen

Ensimmäisessä lähestymistavassa puuttuvat arvot yksinkertaisesti hylätään kaikista laskennoista. Tämä menetelmä on järkevä yllättävän usein, sillä analyysi on yleensä yleistys alkuperäisistä arvoista. Menetelmä on kuitenkin soveltuva vain silloin, kun puuttuvat arvot eivät aiheuta kumulatiivisia vaikutuksia analyysituloksiin. Toisessa lähestymistavassa puuttuvat arvot interpoloidaan esimerkiksi lineaarisella interpolaatiomenetelmällä. Interpolointi voidaan tehdä joko tietokannan sisältöä täydennettäessä tai kyselyiden tuloksia muodostettaessa. Kolmannessa vaihtoehdossa puuttuvat arvot korvataan toisen aikasarjan arvolla, esimerkiksi ennusteella. Neljännessä vaihtoehdossa puuttuvat arvot vaikuttavat suoraan lopputuloksen tarkkuuteen, joka ilmoitetaan erikseen analyysitulosten osana. Tarkkuuden laskentamenetelmä riippuu keräyttämisfunktiosta ja analyysin luonteesta.

Käytännön sovelluksissa OLAP-tekniikka tulisi rakentaa olemassaolevan automaatiojärjestelmän osaksi tai liittää järjestelmän rajapinnan kautta. Nykyisten järjestelmien sulkeutuneisuuden johdosta yleispätevän tietovarastokonseptin laatiminen on vaikeaa etenkin jos tietovarastoon on tarkoitus tallettaa kaikki teollisuusprosessissa oleva tieto. Käytännössä ongelma voidaan ratkaista siten, että analyysimalliin valitaan sopiva kombinaatio prosessimuuttujia ja niiden mittaustietoja kerätään

analyysitarkoitusta varten erilliseen nopeaan aikasarjatietokantaan, jossa datakuutiolaskennat voidaan tehdä optimoidulla tavalla. Yhteenä tähän tarkoitukseen sopivana tietokantaohjelmistona on VTT Tietotekniikassa kehitetty RapidBase[7], jossa käytetään laajennettua relaatiomallia aikasarjojen esittämiseen [WAP99] .

5. LÄHDELUETTELO

[AAD96] Sameet Agarwal, Rakesh Agarwal, Prasad M. Deshpandre, Asnish Gupta, Jeffrey F. Naughton, Ragnu Ramakrishnan, Sunita Sarawagi: On the Computation of Multidimensional Aggregates. Proceedings of the 22nd VLDB Conference, Bombay, India, 1996, sivut 506-521.

[CCS93] F. Codd, S. B. Codd, C. T. Salley: Providing OLAP (On-line Analytical Processing) to User-Analysts: An IT Mandate. Technical report, E.F. Codd & Associates, 1993, http://www.arborsoft.com/essbase/wht_ppr/coddTOC.html

[GBLB96] Jim Gray, Adam Bosworth, Andrew Layman, Hamid Pirahesh: Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. Proceedings of 12th International Conference on Data Engineering, New Orleans, Louisiana, USA, 1996, sivut 152-159.

[Hov97] Ari Hovi: Data Warehousing - tietovarastotekniikka. Suomen Atk-kustannus Oy, Espoo, Suomi, 1997. ISBN 951-762-509-X.

[ISO92] ISO/IEC 9075. Information processing systems - Database language SQL. International standard, third edition, 1992. Ref. No. ISO 9075 : 1992 (E).

[KWPA99] Jukka Kiviniemi, Antoni Wolski, Antti Pesonen, Johannes Arminen: Lazy Aggregates for Real-Time OLAP. Proc. First International Conference on Data Warehousing and Knowledge Discovery (DaWak'99), Aug. 30 - Sep. 1, 1999, Florence, Italy. Lecture Notes in Computer Science, Springer-Verlag, 1999.
http://www.vtt.fi/tte/projects/industrialdb/publs/lazy-aggr.pdf

[WAP99] Antoni Wolski, Johannes Arminen, Antti Pesonen: Relaatiotemporaalinen tietomalli mittaustiedon hallintaa varten. Automaatio 1999, 14-16.9.1999, Helsinki.

[7] http://www.vtt.fi/tte/projects/rapid/