Vesa Pekkarinen

Tietovarastointi, OLAP ja tiedon louhinta

Tietojärjestelmätieteen
kandidaatintutkielma
10.9.2007

Jyväskylän yliopisto
Tietojenkäsittelytieteiden laitos
Jyväskylä

TIIVISTELMÄ

Pekkarinen, Vesa Tapio
Tietovarastointi, OLAP ja tiedon louhinta / Vesa Pekkarinen
Jyväskylä: Jyväskylän yliopisto, 2007, 28s.

Kandidaatintutkielma

Tietovarastointi, OLAP ja tiedon louhinta pyrkivät vastaamaan organisaatioiden liiketoimintatiedon varastointi-, käsittely- ja analysointitarpeisiin. Nämä käsitteet ovat olleet käytössä kirjallisuudessa jo toistakymmentä vuotta mutta niiden merkityksistä on olemassa hyvinkin poikkeavia käsityksiä. Tämän tutkielman tavoitteena on muodostaa yleiskuva tietovarastoinnista, OLAP:ista ja tiedon louhintaan liittyvistä asioista ja tuoda esiin yhteyksiä niiden välillä. Tutkielman avulla pyritään yhdenmukaistamaan kirjallisuudessa termeistä esiintyviä käsityksiä. Tutkielman tutkimusmenetelmänä on kirjallisuuskatsaus.

Tietovarastointi määritellään tutkielmassa prosessiksi, jonka tavoitteena on tukea päätöksentekoa keräämällä tietoa tietolähteistä tietovarastoon, ja jakamalla sitä asiakassovellusten käytettäväksi. OLAP määritellään lähestymistavaksi, joka määrittää miten tietoa tulee varastoida, käsitellä ja visualisoida, jotta sitä voidaan hyödyntää päätöksenteossa. Tiedon louhinta puolestaan määritellään tekniikaksi, jota käytetään löytämään uutta tietämystä suurista tietomääristä, ja tämän tietämyksen esittämistä kaavoina, malleina tai sääntöinä. Tiedon louhinta on osa prosessia, jota kutsutaan tietämyksen löytämiseksi tietokannasta.

Tietovarastointi ja OLAP ovat hyvin kiinteästi toisiinsa liittyviä käsitteitä. OLAP:n yhteydessä käytetään usein hyödyksi organisaation tietovarastoja. Myös tiedon louhinta tehostuu kun louhitaan tietoa tietovarastoista operatiivisten tietokantojen sijaan. On olemassa myös tutkimusta OLAP:n ja tiedon louhinnan yhdistämisestä, minkä tarkoituksena on auttaa päätöksentekijöitä analysoimaan louhittua tietämystä tulosten visualisoinnin avulla.

AVAINSANAT: Tietovarastointi, Data Warehousing, OLAP, tiedon louhinta, data mining

Ohjaaja: Mauri Leppänen
Tietojenkäsittelytieteiden laitos
Jyväskylän Yliopisto

Tarkastaja: Jorma Kyppö
Tietojenkäsittelytieteiden laitos
Jyväskylän Yliopisto

SISÄLLYSLUETTELO

1 JOHDANTO

Jokapäiväisessä toiminnassaan organisaatiot tuottavat suuria määriä tietoa liiketoimintaansa liittyvistä tapahtumista. Tällaista tietoa ovat esimerkiksi vähittäiskauppojen myyntitieto, pankkitileihin kohdistuneet tilitapahtumat ja asiakaspalveluhenkilöstön kirjaamat palvelutoimenpiteet. Tätä kertyvää tietoa pyritään hyödyntämään organisaatioissa monin tavoin päätöksenteon tukena. Yritysjohto pyrkii ennustamaan historiatiedon avulla asiakkaiden käyttäytymisen trendejä. Sisäänostajat puolestaan tutkivat varastotilanteita ja myyntilukuja ja päättävät niiden perusteella, mitä tuotteita tilataan tukkukauppiaalta ja mitkä tuotteet myydään alennuksella.

On kehitetty useita lähestymistapoja, menetelmiä ja niitä tukevia teknologioita, jotka pyrkivät vastaamaan tämän tiedon käsittely- ja analysointitarpeeseen omalla tavallaan. Tässä tutkielmassa käsitellään näistä pääasiallisesti seuraavia: tietovarastointi (DW, data warehousing), OLAP (online analytical processing) sekä tiedon louhinta (DM, data mining). Edellä mainittuihin liittyy myös läheisesti päätöksenteon tukijärjestelmän käsite (DSS, decision support system).

Päätöksenteon tueksi tarvitaan nykyisen tiedon lisäksi historiallista tietoa eri organisaation tietolähteistä (data source) koottuna. Tietovarastot on kehitetty mahdollistamaan tällainen tiedon analysointi. Tieto kerätään tietolähteistä tietovarastoon (data warehouse), josta sitä voidaan hyödyntää analysointiin käytettävillä teknologioilla. Näistä kaksi tärkeintä ovat OLAP-työkalut ja tiedon louhinta. Liiketoimintatiedon hallinnaksi (BI, business intelligence) kutsutaan kokonaisuutta, jossa OLAP:ia, tiedon louhintaa tai näitä molempia käytetään tietovarastointiin yhdistettynä. (Connolly & Begg 2005, 1204)

OLAP on termi, jota käytetään kuvaamaan tietovarastojen monimutkaisen datan analysoimista (Elmasri & Navathe 2007, 978). McFaddenin, Hofferin ja Prescottin (1999, 560) mukaan OLAP on graafisten työkalujen käyttämistä. Nämä työkalut tarjoavat käyttäjille dataan moniulotteisia näkymiä, joiden kautta käyttäjät voivat analysoida tietoa käyttämällä yksinkertaisia ikkunointitekniikoita.

Tiedon louhinta (DM, Data Mining) voidaan nähdä yhtenä osana laajempaa prosessia, josta käytetään termiä tietämyksen löytäminen tietokannasta (KDD, Knowledge Discovery in Databases). Tiedon louhinnan tavoitteena voi olla esimerkiksi asiakkaiden ostokäyttäytymisen ennustaminen tai mielenkiintoisten ilmiöiden tunnistaminen tietokantaan tallennetun tiedon perusteella. Tähän pyritään louhimalla tarjolla olevasta datasta tietämystä, joka voidaan esittää muiden muassa sääntöinä, kaavoina, päätöspuina tai semanttisina verkkoina. (Elmasri & Navathe 2007, 946-948)

Vaikka yllä mainitut käsitteet ovat olleet käytössä jo toistakymmentä vuotta, niiden merkityksestä on olemassa hyvinkin poikkeavia käsityksiä. Tämän tutkielman tavoitteena on muodostaa yleiskuva tietovarastointiin, OLAP:iin ja tiedon louhintaan liittyvistä asioista. Tutkielmassa pyritään yhdenmukaistamaan keskeisille termeille kirjallisuudessa annettuja käsityksiä.

Luku kaksi käsittelee tietovarastointia, sen käyttötarvetta, yleisimpiä tietovarastoarkkitehtuurimalleja sekä tiedon tallennusmalleja. Luvussa kolme tarkastellaan OLAP – käsitettä ja teknologiaa. Käsiteltävinä asioina ovat OLAP:n käyttötarkoitus ja sen tarjoamat operaatiot tiedon analyysin tueksi. Luku neljä selvittää mitä tiedon louhinta on, mihin sitä tarvitaan ja millaisin keinoin tietoa louhitaan. Luvussa viisi esitetään yhteenveto, jossa tuodaan esiin yhteyksiä näiden kolmen käsitteen välillä.

2 TIETOVARASTOINTI

Tietovarastointi on vahvasti yhteydessä seuraavissa luvuissa käsiteltäviin OLAP:iin ja tiedon louhintaan. Vaikka tietovarastointi on kehittynyt ennen OLAP:ia, sitoo kirjallisuus usein nämä kaksi aihepiiriä hyvin tiukasti toisiinsa. OLAP nähdään useimmiten tietovarastoa käyttävänä asiakkaana (client) mutta sen vaikutus näkyy myös palvelinpuolella (server) selvästi, esimerkiksi palvelintyyppien ja tietomallien kehittymisessä ja termistössä. Nämä aiheet käsitellään kuitenkin tässä luvussa, koska ne liittyvät OLAP:n lisäksi myös muihin tietovarastoja hyödyntäviin teknologioihin, kuten tiedon louhintaan. OLAP:ia ja tiedon louhintaa käsittelevissä luvuissa 3 ja 4 kerrotaan myös lisää niiden yhteyksistä tietovarastointiin.

Tässä luvussa esitetään ensin määritelmä ja yleinen kuvaus tietovarastoinnista, sekä kerrotaan mihin tietovarastointia tarvitaan. Seuraavissa kohdissa kuvataan yleisimpiä tietovarastoarkkitehtuureja sekä tietovarastoissa käytettyjä tiedon tallennusmalleja.

2.1 Tietovaraston ja tietovarastoinnin käsitteet

Koska tietovarastoja on kehitetty useissa organisaatioissa vastaamaan tiettyihin tarpeisiin, ei termille tietovarasto ole olemassa yksittäistä tunnustettua määritelmää (Elmasri & Navathe 2007, 978). Useissa lähteissä (mm. Chaudhuri & Dayal 1997; Connolly & Begg 2005; Elmasri & Navathe 2007; McFadden, Hoffer & Prescott 1999) viitataan kuitenkin William Inmonin esittämään määritelmään. Inmonia on pidetty tietovarastoinnin isänä hänen käytettyään ensimmäisenä käsitettä ”warehouse” tietovarastosta puhuttaessa (Connolly & Begg 2005, 1151; Elmasri & Navathe 2007, 978). Inmonin (1996) määritelmän mukaan tietovarasto on ”a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management’s decision-making process”. Inmon siis määrittelee tietovaraston kokoelmaksi päätöksentekoprosessin tukena käytettävää tietoa. Tämä tieto on siis tietyiltä aihealueilta kerättyä, integroitua, aikasidonnaista ja tietovarastoon lataamisen jälkeen muuttumatonta. Nämä ominaisuudet erottavat tietovarastoidun tiedon operatiivisten tietokantojen tiedosta.

Myöskään tietovarastoinnille ei löydy vakiintunutta määritelmää. Chaudhuri ja Dayal (1997) määrittelevät tietovarastoinnin kokoelmaksi päätöksenteon tukemiseen tarkoitettuja teknologioita, joilla pyritään auttamaan tietämystyöntekijöitä tekemään parempia ja nopeampia päätöksiä. Barquinin (1996) määritelmän mukaan McFadden, Hoffer ja Prescott (1999, 531) esittävät tietovarastoinnin prosessina, jonka avulla organisaatiot poimivat (extract) sisältöä informaatiovarannostaan (informational assets) käyttämällä hyväkseen tietovarastoja.

Tässä tutkielmassa tietovaraston määritelmänä käytetään Inmonin (1996) määritelmää. Tietovarastointi määritellään prosessiksi, jonka tavoitteena on tukea päätöksentekoa keräämällä tietoa tietolähteistä tietovarastoon, ja jakamalla sitä asiakassovellusten käytettäväksi. Prosessi koostuu viidestä tietovirrasta, joiden mukaan tietoa siirretään ja muokataan (kts. kohta 2.3).

2.2 Tarve tietovarastoinnille

Tiedon analysoinnin välineiden ja tekniikoiden kasvava kehittyminen on johtanut siihen, että tarvitaan kykyä tallentaa valtavia tietomääriä. Lisäksi tarvitaan toiminnallisuutta ja tehoa vastata tähän tietomäärään kohdistuviin monimutkaisiin kyselyihin. Nämä vaatimukset ylittävät perinteisten tapahtumien hallintaan tarkoitettujen tietokantajärjestelmien (OLTP, online transaction processing) kyvyt. (Elmasri & Navathe 2007, 977)

Tietovarastoinnin kehittymiseen ja yleistymiseen on johtanut pääasiassa kaksi päätekijää: (1) tieto yritysten operatiivisissa tietokannoissa on tyypillisesti pirstoutunutta ja huonolaatuista ja usein hajallaan erilaisilla yhteensopimattomilla laitteisto- ja sovellusalustoilla ja (2) tarve erottaa tietokeskeiset järjestelmät (informational systems) operatiivisista järjestelmistä (operational systems), koska niiden perusominaisuudet ja käyttötarpeet eroavat toisistaan huomattavasti. (McFadden, Hoffer & Prescott 1999, 532-534) TAULUKKO 1 esittää operatiivisten tapahtuman hallintajärjestelmien ja tietovarastojärjestelmien eroja.

Operatiiviset tietokantajärjestelmät on suunniteltu käsittelemään suuria määriä jäsennettyjä ja toistuvia atomisia tapahtumia (transactions). Nämä tapahtumat vaativat tarkkaa, ajantasaista tietoa, ja ne lukevat tai päivittävät yleensä korkeintaan muutamia kymmeniä tietueita. Operatiivisten tietokantojen koko on tyypillisesti sadoista gigatavuista muutamiin teratavuihin. Eheys ja palautettavuus ovat näiden järjestelmien kriittisiä ominaisuuksia. (Chaudhuri & Dayal 1997)

Tietovarastot sen sijaan on tarkoitettu päätöksenteon tukemiseen. Historian kattava ja yhdistetty summatieto on tärkeämpää kuin tarkat, yksittäiset tietueet. Tietovarastot sisältävät yhdistettyä tietoa mahdollisesti useista operatiivisista tietokannoista, joten ne ovat usein niitä kokoluokkaa suurempia. Yrityksen tietovarastot voivat olla kooltaan jopa satoja teratavuja. Työtaakka koostuu enimmäkseen monimutkaisista ennakoimattomista kyselyistä, jotka käsittelevät miljoonia tietueita ja suorittavat paljon selaushakuja, liitoksia ja koostamista. Tietovarastoille tärkeitä ominaisuuksia ovat siis suoritusteho kyselyjen vaste-ajan suhteen. (Chaudhuri & Dayal 1997) Jos kyselyt kohdistettaisiin suoraan operatiivisiin tietokantoihin tietovaraston sijaan, saattaisi järjestelmä kuormittua liikaa ja sen varsinainen toiminta vaarantua.

Taulukko 1. Tapahtuman hallintajärjestelmien ja tietovarastojärjestelmien vertailu (Connolly & Begg 2005, 1153).

Tapahtuman hallintajärjestelmät (OLTP systems)	*Tietovarastojärjestelmät (data warehousing systems)*
Sisältää nykyistä tietoa	Sisältää historiallista tietoa
Tallentaa yksityiskohtaista tietoa	Tallentaa yksityiskohtaista, kevyesti koostettua ja vahvasti koostettua tietoa
Tieto on dynaamista	Tieto on enimmäkseen staattista
Tiedon käsittely on toistuvaluonteista	Tiedon käsittely on ennakoimatonta, suunnittelematonta ja heuristista
Tietokantatapahtumien suoritusmäärät suuret	Tietokantatapahtumien suoritusmäärät alhaiset tai keskiverrot
Järjestelmän käyttäjän toimintamalli ennustettava	Järjestelmän käyttäjän toimintamalli ennustamaton
Tapahtumaherätteinen	Analyysiherätteinen
Tiettyihin sovelluksiin suuntautunut	Tiettyihin aiheisiin suuntautunut
Tukee päivittäisiä päätöksiä	Tukee strategisia päätöksiä
Palvelee suurta määrää operatiivisia käyttäjiä	Palvelee suhteellisen pientä määrää johtajatason käyttäjiä

2.3 Tietovarastointiprosessi

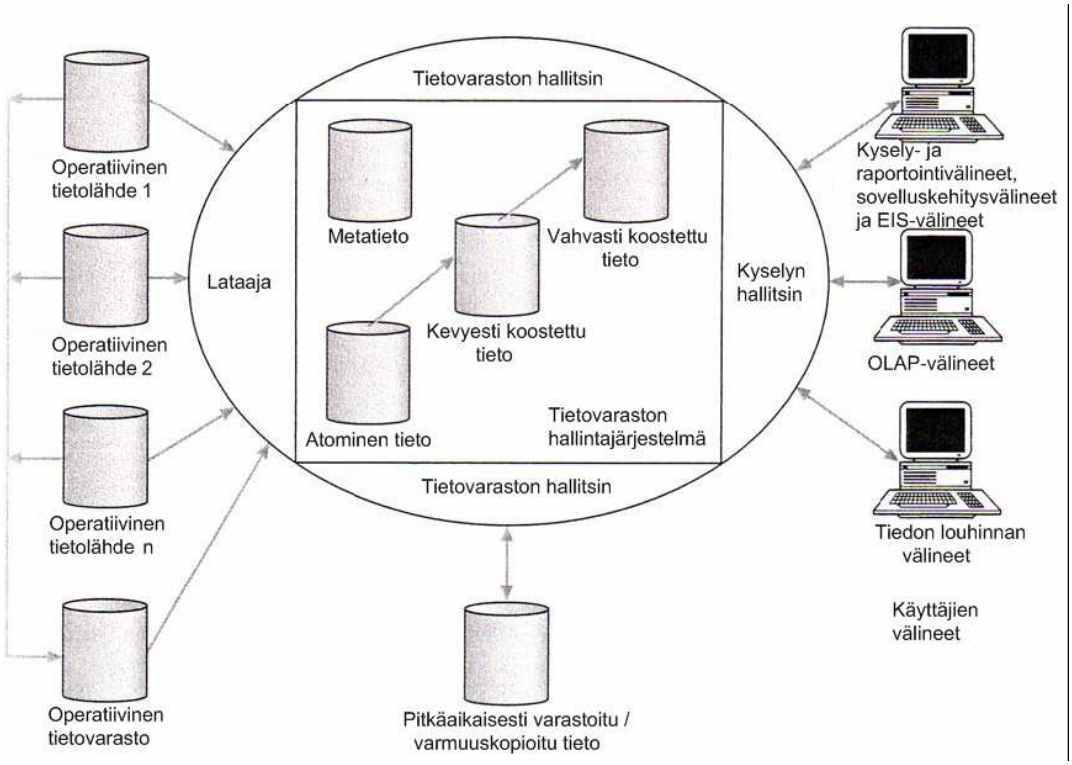
Ennen kuin tieto on valmiina hyödynnettäväksi tietovarastossa, on sitä käsiteltävä. Tätä tietovarastoinnin tietovirtaa kutsutaan sisäänvirtaukseksi (inflow) (Connolly & Begg 2005, 1162). Elmasri ja Navathe (2007, 986) ovat listanneet tähän tietovarastointiprosessin osaan seuraavat vaiheet: (1) tiedon poiminta tietolähteistä (extracting), (2) tiedon muotoilu yhtenäiseksi (formatting), (3) tiedon puhdistus (cleaning), (4) tiedon sovitus tietovaraston tietomalliin (fitting) ja (5) tiedon lataaminen (loading).

Jotta monimutkaisten, usein ennakoimattomien, kyselyiden tekeminen ja nopea suorittaminen olisi mahdollista, tulee tietovarastojen tarjota paljon laajempi ja tehokkaampi kyselytuki kuin tapahtuman hallintajärjestelmien. Tietovarastot tarjoavat tuen varastoidun tiedon hyödyntämiseen mm. kehittyneille taulukkolaskentaohjelmille, OLAP-sovelluksille ja tiedon louhinnan työkaluille (Elmasri & Navathe 2007, 988). Tätä käyttäjille päin näkyvää tietovirtaa kutsutaan ulosvirtaukseksi (outflow) (Connolly & Begg 2005, 1162).

Sisäänvirtauksen ja ulosvirtauksen lisäksi on määritelty kolme muuta tietovirtaa: ylösvirtaus (upflow), alasvirtaus (downflow) ja metavirtaus (metaflow) (Connolly & Begg 2005, 1162). Ylösvirtauksessa tiedosta saatavaa hyötyä parannetaan esimerkiksi koostamalla sitä käyttäjille hyödyllisempään muotoon ja hajauttamalla sitä eri palvelimille. Alasvirtaus tarkoittaa vanhan tiedon pitkäaikaisempaa varastointia ja varmuuskopiointien suorittamista. Tietovarastojen rakentamista, ylläpitoa ja käyttöä varten tarvitaan myös paljon metatietoa, jonka hallintaa kutsutaan metavirtaukseksi.

2.4 Arkkitehtuurimallit

Organisaation eri käyttäjäryhmien, esimerkiksi yrityksen osastojen, tarpeista riippuen tietovaraston arkkitehtuuri ja tallennetun tiedon esittämisen tarkkuus voi vaihdella. Kirjallisuuslähteissä esitetään hieman toisistaan eroavia arkkitehtuurimalleja tietovarastolle (kts. esimerkiksi Connolly & Begg 2005, 1156; Jarke, Lenzerini, Vassiliou & Vassiliadis 1999, 10-12). Niistä voidaan kuitenkin löytää sama perusajatus. Tieto kerätään tietolähteistä organisaation tietovarastoon joko suoraan tai operatiivisen tietovaraston kautta (ODS, operational data store). Kerätty tieto puolestaan jaetaan käyttäjien saataville, yleensä joko paikallisvarastoihin (data mart) tai erityisille OLAP-palvelimille. KUVIO 1 esittää tyypillistä tietovarastoarkkitehtuuria.



Kuvio 1. Tyypillinen tietovarastoarkkitehtuuri (Connolly & Begg 2005, 1157).

Esimerkkinä arkkitehtuurimalleista tässä esitellään Jarken ym. (1999, 10-12) mukaisesti kolme perusmallia tietovaraston fyysiselle arkkitehtuurille: keskitetty arkkitehtuuri (centralized architecture), yhdistetty arkkitehtuuri (federated architecture) sekä kerrosarkkitehtuuri (tiered architecture).

Keskitetyssä arkkitehtuurissa käytetään vain yhtä tietovarastoa, johon kerätään organisaation tietolähteistä kaikki tarvittava tieto analyysia varten. Tietoa hyödyntävät käyttäjät ja järjestelmät tekevät kyselyjä suoraan tähän keskitettyyn tietovarastoon.

Yhdistetty arkkitehtuuri pyrkii mukailemaan organisaation loogista rakennetta tiedon hyödyntämisen suhteen. Tieto on loogisesti yhdistettyä, mutta fyysisesti tallennettu erillisiin paikallisvarastoihin. Paikallisvarastot palvelevat eri käyttäjäryhmien tarpeita, esimerkiksi organisaation eri osastoja ja tieto on karsittu käyttäjäryhmien mukaan. Koska tietoa on paikallisvarastoissa vähemmän kuin keskitetyssä mallissa, sitä voidaan varastoida kaikissa tarkkuusasteissa atomisesta hyvin koosteiseen.

Kerrosarkkitehtuurissa sen sijaan käytetään yhteistä fyysistä tietovarastoa. Atomista tietoa varastoidaan vain keskustietovarastossa, josta sitä koostetaan ja kopioidaan eteenpäin paikallisvarastoihin. Näitä voi olla yhdessä tai useammassa kerroksessa ennen loppukäyttäjiä. Myös tiedon kerääminen tietolähteistä voi tapahtua kerroksittain, esimerkiksi ensin myymäläkohtaisesti, sitten alueittain ja vasta lopuksi organisaation maanlaajuiseen tietovarastoon.

2.5 Palvelintyyppejä

Tietovarastojen palvelinalustoina voidaan käyttää tavallisia relaatiotietokantapalvelimia, jotka toimivat hyvin varsinaisessa atomisen tiedon varastoinnissa. Ne eivät kuitenkaan tarjoa riittäviä ominaisuuksia, joiden avulla voidaan tehokkaasti muodostaa päätöksenteon tukijärjestelmien tarvitsemia moniulotteisia näkymiä tietoon (Chaudhuri & Dayal 1997). Chaudhurin ja Dayalin mukaan tähän tarkoitukseen on olemassa kolmeen kategoriaan luokiteltavia palvelimia: erikoistuneet SQL-palvelimet, ROLAP-palvelimet ja MOLAP-palvelimet.

Erikoistuneet SQL-palvelimet ovat relaatiotietokantapalvelimia, jotka tarjoavat tukea SQL-kyselyille ja moniulotteisten tietomallien toteuttamiseen (Chaudhuri & Dayal 1997). Esimerkiksi Red Brick Warehouse on tällainen erikoistunut palvelin, ja se tarjoaa oman toimittajaspesifin SQL-kielen nimeltään RISQL tietovaraston käyttämiseen.

ROLAP-palvelimet (relational OLAP) toimivat välikerroksena tiedon varsinaisesti tallentavien relaatiotietokantapalvelimien ja asiakassovellusten välissä. Tieto ROLAP-palvelimilla tallennetaan relaatioihin, ja palvelimet

tukevat SQL:n laajennuksia ja muita erityisiä tiedon käsittelymenetelmiä moniulotteisen tietomallin toteuttamiseen.

MOLAP-palvelimet (multidimensional OLAP) tallentavat tietoa suoraan erityisiin moniulotteisiin rakenteisiin, kuten taulukoihin. Ne toteuttavat OLAP-operaatioita hyödyntäen suoraan näiden tietorakenteiden moniulotteisuutta (Chaudhuri & Dayal, 1997).

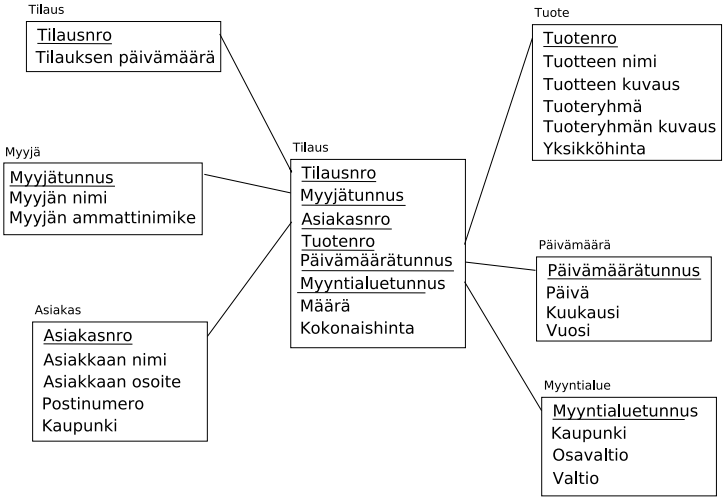
Näiden kolmen kategorian lisäksi on myös eroteltu omaksi ryhmäkseen HOLAP (hybrid OLAP) sekä DOLAP (desktop OLAP). HOLAP yhdistää ROLAP ja MOLAP – palvelinten ominaisuuksia. DOLAP puolestaan yrittää hyödyntää muita laajemmin pöytätietokoneiden laskentatehoa moniulotteisen laskennan suorittamisessa. (Connolly & Begg 2005, 1215-1256)

2.6 Moniulotteiset tietomallit

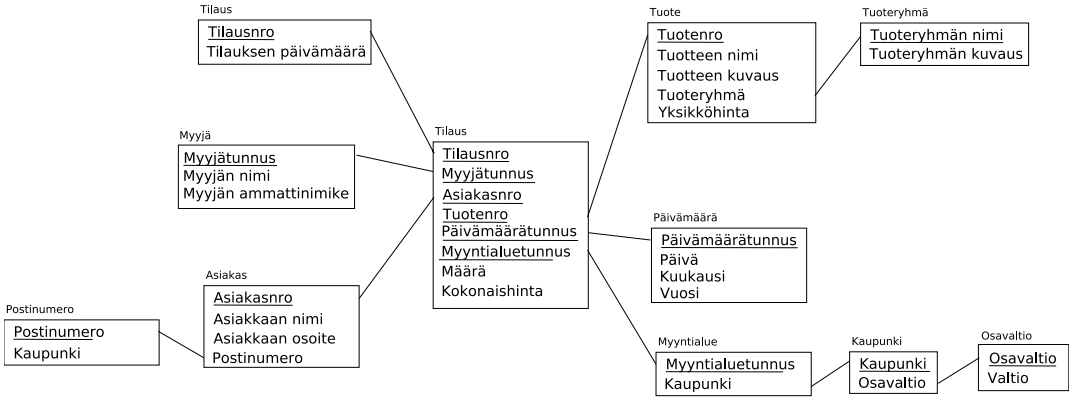
Mahdollistaakseen monimutkaisen analyysin ja visualisoinnin tietovarastot käyttävät tiedon tallentamiseen yleensä moniulotteista tietomallia (Chaudhuri & Dayal 1997). Moniulotteiset tietorakenteet tukevat hyvin OLAP-teknologian ja päätöksenteon tukijärjestelmien tarpeita (Elmasri & Navathe 2007, 979). Myös tiedon louhintaa saadaan tehostettua, kun sen kohteena on koostettu tieto operatiivisten tietokantojen yksittäisten tapahtumien sijaan (Elmasri & Navathe 2007, 946).

Yleisimmät tietovarastoissa käytetyt moniulotteiset mallityypit ovat tähtimalli (star schema) ja lumihiutalemalli (snowflake schema) (Elmasri & Navathe 2007, 983-984). Molemmissa malleissa käytetään kahdenlaisia tauluja. Mallin keskeinen osa on faktataulu (fact table), joka sisältää tietoa tietyistä organisaation liiketoimintatapahtumista, kuten esimerkiksi myynneistä. Faktataulussa voi olla satoja miljoonia rivejä, ja ne muodostavat pääosan tietovarastoon tallennettavasta tiedosta. Faktataulun rivi sisältää tarvittavat laskennalliset tiedot tapahtumasta (esimerkiksi myyntihinta) sekä viiteavaimia ulottuvuustauluihin (dimension table). Ulottuvuustauluihin tallennetaan tietoja, joilla yksittäinen tapahtumatieto voidaan identifioida. Ulottuvuustaulu voi sisältää esimerkiksi myyntitapahtuman aikaleiman tai myynnin suorittaneen osaston tiedot.

Tähtimalli ja lumihiutalemalli eroavat toisistaan ulottuvuustaulujen osalta. Tähtimallissa ulottuvuustaulut ovat normalisoimattomia, joten jokaista tapahtumariviin liittyvää ulottuvuutta kohti on vain yksi taulu (kts. KUVIO 2). Lumihiutalemallissa ulottuvuustaulut on ryhmitelty hierarkisesti normalisoimalla ne (kts. KUVIO 3). Lumihiutalemallin etuja tähtimalliin verrattuna ovat pienempi tilankäyttö ja joustavuus. Suurimpana haittana puolestaan on heikompi suorituskyky (McFadden, Hoffer & Prescott, 1999, 557).



Kuvio 2. Esimerkki tähtimallista.



Kuvio 3. Esimerkki lumihiutalemallista.

MOLAP-palvelimet käyttävät tiedon tallennukseen omia moniulotteisia tallennusmalleja, jotka eivät perustu relaatiomalliin. Useimmiten ne käyttävät jonkinlaisia tähtimallin kaltaisia taulukoita. Näiden mallien hyvänä puolena on nopea toiminta silloin kun suoritetaan kyselyitä, joita varten tietovarasto on optimoitu. Relaatiopohjaisiin tietomalleihin verrattuna nämä moniulotteiset mallit vievät kuitenkin huomattavasti enemmän tallennustilaa, joten niihin tallennettavan tiedon määrä on rajoitetumpi. (McFadden, Hoffer & Prescott, 1999, 557)

3 OLAP

Edellisessä luvussa käsiteltiin tietovarastointia, jossa mahdollisesti koko organisaation laajuudelta kerätään ja yhdistetään tietoa käytettäväksi päätöksenteon tukena. Tämän tiedon analysointiin on kehitetty teknologioita ja työkaluja, joista tärkeimmät ovat OLAP ja tiedon louhinta (Connolly & Begg 2005, 1150). Tässä luvussa esitellään näistä OLAP.

Ensin määritellään OLAP ja kuvataan sen käyttötarkoitusta. Sen jälkeen esitellään OLAP-järjestelmän keskeiset ominaisuudet ja yleisimmät OLAP-operaatiot.

3.1 OLAP – käsite

Käsitteelle OLAP löytyy kirjallisuudesta useita määritelmiä. Seuraavissa esitetään joitain poimintoja näistä.

Coddin, Coddin ja Salleyn (1993) mukaan OLAP tarkoittaa dynaamista analyysia, jota tarvitaan luomaan, manipuloimaan, elävöittämään ja yhdistämään tietoa yrityksen suurista tietomääristä.

McFadden, Hoffer ja Presscot (1999, 560) määrittelevät OLAP-käsitteen seuraavasti. OLAP on graafisten työkalujen joukon käyttämistä. Nämä työkalut tarjoavat käyttäjille moniulotteisia näkymiä dataan, joiden kautta käyttäjät voivat analysoida tietoa käyttämällä yksinkertaisia ikkunointitekniikoita.

Elmasrin ja Navathen (2007, 978) mukaan OLAP on termi, jota käytetään kuvaamaan tietovaraston monimutkaisen datan analysoimista. Taitavien tietämystyöntekijöiden (knowledge worker) käsissä OLAP-työkalut tarjoavat hajautetun laskennan mahdollisuuksia analyyseissa, jotka vaativat enemmän tallennustilaa ja suoritustehoa kuin yksittäiseen pöytätietokoneeseen voidaan sijoittaa taloudellisesti ja tehokkaasti.

OLAP:n avulla voidaan analysoida tietovarastoihin tallennettua tietoa. Sitä voidaan hyödyntää organisaatioissa monin eri tavoin. Esimerkkejä OLAP:n käyttötavoista ovat mm. budjetointi, taloudellisten mallien tekeminen, myynnin analysoiminen ja ennustaminen, markkinatutkimus, asiakasanalyysi ja asiakkaiden segmentointi (OLAP Council 1997). Tietovarastoista voidaan hakea vastauksia kysymyksiin ”kuka?” ja ”mitä?” käyttämällä esimerkiksi yksinkertaisia raportointisovelluksia. OLAP-työkaluilla voidaan suorittaa tietovarastoihin ja OLAP-palvelimille monimutkaisempia kyselyitä, joiden avulla voidaan vastata myös kysymyksiin ”mitä jos?” ja ”miksi?”. (Connolly & Begg 2005, 1205)

OLAP ja tietovarastot ovat toisiaan täydentäviä menetelmiä. Tietovarasto tallentaa ja käsittelee tietoa, ja OLAP muuttaa sen strategiseksi informaatioksi. OLAP:n tarjoamat toiminnot ulottuvat tietovarastoidun tiedon tutkimisesta navigoimalla ja selaamalla aina monimutkaisten mallien ja aikasarjojen analysointiin. (OLAP Council 1997) OLAP voi toimia myös itsenäisenä, organisaation varsinaisesta tietovarastosta riippumattomana järjestelmänä. Ennen kuin itsenäiselle OLAP-palvelimelle tallennettu tieto on tehokkaasti hyödynnettävissä, on sille suoritettava samat toimenpiteet kuin tietovarastoidulle tiedolle. OLAP:n tietovirta on siis suurelta osin samankaltainen tietovarastoinnin kanssa. Jos organisaatiossa on olemassa oleva tietovarasto, on sitä luontevaa käyttää OLAP:n perustana.

OLAP:ia käsittelevä kirjallisuus käsittelee tiedon varastointiin, käsittelyyn ja analysointiin liittyviä prosesseja sekä niihin liittyviä teknologioita. Kirjallisuudessa tarkastellaan millaisessa muodossa tietoa tallennetaan, millaisia ominaisuuksia OLAP-palvelimilla ja –työkalusovelluksilla on, ja millaisia toiminnallisuuksia niiden tulee tarjota, jotta niiden avulla voidaan suorittaa analyyseja suuriin tietomääriin. Tässä tutkielmassa OLAP määritelläänkin lähestymistavaksi, joka määrittää miten tietoa tulee varastoida, käsitellä ja visualisoida, jotta sitä voidaan hyödyntää päätöksenteossa.

3.2 Keskeiset ominaisuudet

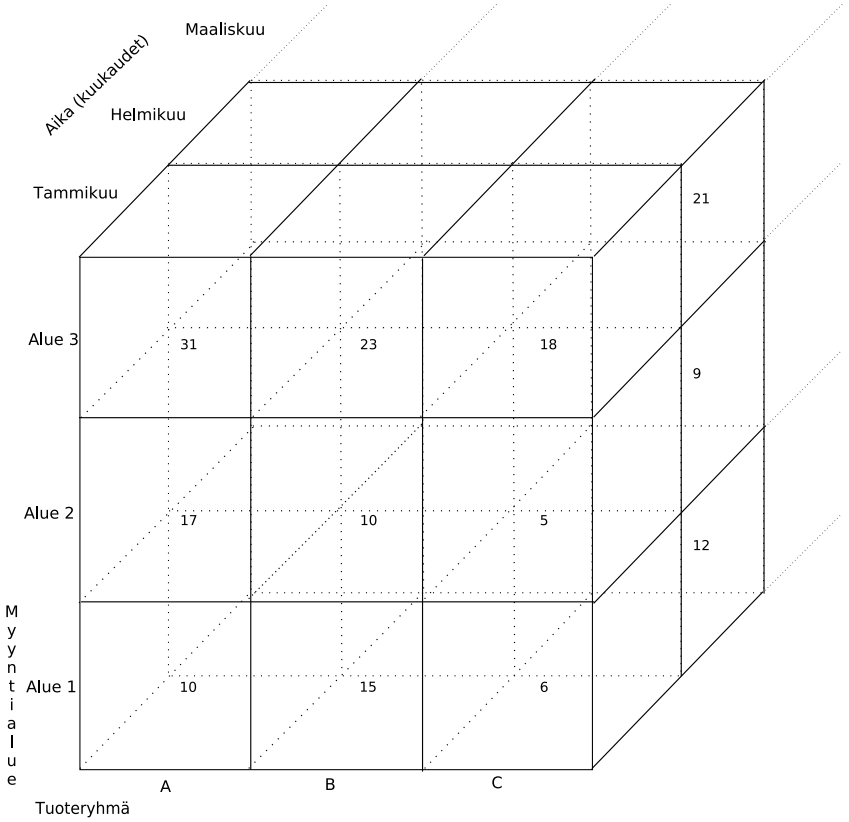
OLAP Council (1997) on määritellyt kolme keskeistä ominaisuutta, jotka OLAP-järjestelmien tulisi toteuttaa: moniulotteiset näkymät tietoon (multidimensional views of data), kyky suorittaa monimutkaisia laskutoimituksia (calculation-intensive capabilities) sekä älykäs ajan käsittely (time intelligence). Toinen tunnettu OLAP-tuotteiden arviointiin tarkoitettu säännöstö on Coddin 12 sääntöä (Codd, Codd & Salley, 1993). Seuraavissa kappaleissa selitetään tarkemmin edellä mainitut OLAP Councilin esittämää kolme ominaisuutta.

3.2.1 Moniulotteinen näkymä tietoon

Moniulotteiset näkymät ovat käyttökelpoisia esitysmuotoja analysoidessa organisaation tietovarastoja. Esimerkiksi OLAP-kysely voisi olla ”Kuinka monta tuoteryhmien A, B ja C tuotetta myytiin eri myyntialueilla, vuoden 2003 eri kuukausina?”. Tällaisen kyselyn tulosten esittäminen kaksiulotteisena tulostauluna ei ole kovin havainnollinen tapa. Moniulotteiset rakenteet pystytään esittämään parhaiten tietokuutioina (data cube), joissa jokainen sivu vastaa yhtä ulottuvuutta (Connolly & Begg 2005, 1209).

KUVIO 4 esittää tietokuutiota, joka voisi muodostua edellä mainitun kyselyn seurauksena. Kuution solujen sisältö kertoo tuotteiden myyntimäärät, ja kuution ulottuvuuksina ovat tuoteryhmät, myyntialueet ja kuukaudet. Kyselyn

tuloksena muodostettua tietokuutiota voidaan edelleen käsitellä ja analysoida OLAP-operaatioiden avulla. Ulottuvuudet ovat usein luonteeltaan hierarkkisia, ja kuutiossa voidaan esimerkiksi kuukausien suhteen porautua (drill-down) tarkemmalle tasolle tarkastelemaan tuloksia viikoittain, ja edelleen päivittäin. OLAP-operaatioita on kuvataan tarkemmin kohdassa 3.3.



Kuvio 4. Esimerkki myyntimääriä kuvaavasta tietokuutiosta, jonka ulottuvuuksina ovat tuoteryhmä, myyntialue ja aika (kuukaudet).

OLAP-järjestelmän tulee siis tarjota moniulotteisten näkymien muodostamiseen tarvittavat operaatiot. Lisäksi järjestelmällä on oltava riittävä laskentateho moniulotteisiin rakenteisiin kohdistuvien kyselyjen tekemiseen ja tulosten käsittelyyn OLAP-operaatioilla. Tätä ominaisuutta käsitellään seuraavaksi.

3.2.2 Monimutkaiset laskutoimitukset

Monimutkaiset OLAP-kyselyt saattavat vaatia paljon arvojen hakemista tietokannasta, ja näiden hakutulosten koostamista. Yksinkertaisten koostefunktioiden, kuten summa tai keskiarvo, lisäksi OLAP-järjestelmien tulee kyetä suorittamaan tehokkaasti myös monimutkaisempaa laskentaa. Tällaista

ovat esimerkiksi liikkuvien keskiarvojen ja prosenttiosuuksien (share calculations) laskeminen. (OLAP Council 1997)

Kuten kohdassa 2.6 kerrottiin, tietovarastot ja varsinkin OLAP-palvelimet käyttävät moniulotteista tietomallia tiedon tallentamiseen. Tämä mahdollistaa sen, että käyttäjä pystyy suorittamaan OLAP-operaatioita lähes reaaliaikaisesti tietokuutiota käsitellessään. Tyypillinen relaatiotietokannan hallintajärjestelmä voi käsitellä satoja tietueita sekunnissa, kun taas tyypillinen moniulotteinen tietokannan hallintajärjestelmä (multi-dimensional DBMS) voi suorittaa 10000 tai useamman tietueen koostamisen sekunnissa (Connolly & Begg 2005, 1209).

Ulottuvuuksien lisääminen moniulotteiseen kyselyyn vaikuttaa käsiteltävien tietueiden määrään eksponentiaalisesti. Laskennan nopeuttamiseksi tietoa tallennetaan OLAP-palvelimille osittain valmiiksi esikoostettuna (pre-aggregate, consolidate), jolloin jokaista arvoa ei jouduta laskemaan uudestaan. Tämä on erityisen hyödyllistä, koska ulottuvuudet ovat yleensä hierarkkisia ja esikoostaminen mahdollistaa alemmalle ulottuvuuden tasolle porautumisen (kts. kohta 3.3). (Connolly & Begg 2005, 1209) Esikoostaminen nopeuttaa laskentaa huomattavasti, mutta vaatii enemmän tallennuskapasiteettia.

3.2.3 Älykäs ajan käsittely

Aika on tärkeä käsite tietovarastoinnin ja OLAP:n yhteydessä. Kuten Inmonin (1996) OLAP-määritelmä ilmaisee, tietovarastoihin tallennettu tieto sisältää aina aikaulottuvuuden. Aikaulottuvuus on myös lähes aina mukana yrityksen toimintaa koskevissa analyyseissa (OLAP Council 1997). Älykäs ajan käsittely on yksi OLAP Councilin (1997) esittämistä vaatimuksista hyvälle OLAP-järjestelmälle. Tämä tarkoittaa OLAP Councilin mukaan sitä, että järjestelmä ymmärtää ajan peräkkäisen luonteen ja sen avulla tulee myös pystyä helposti määrittelemään aikaan liittyviä käsitteitä, kuten aika vuoden alusta tähän päivään (year-to-date). OLAP Councilin mukaan OLAP-järjestelmän tulisi myös kyetä suorittamaan erilaisia laskutoimituksia ajan suhteen, kuten tase ajanjakson aikana (balances over time).

3.3 OLAP-operaatiot

Kohdassa 3.2.1 kerrottiin, että OLAP-järjestelmät esittävät tietoa tietokuutioina. Tapahtumasarjaa, jossa käyttäjä tarkastelee tietokuutiota interaktiivisesti OLAP-työkalusovelluksen tarjoamilla operaatioilla, kutsutaan navigoimiseksi (Jarke, Lenzerini, Vassiliou & Vassiliadis 2000, 90). Seuraavassa kuvaillaan yleisimmin käytetyt OLAP-operaatiot Chaudhurin ja Dayalin (1997) sekä Jarken ym. (2000) mukaan. Käytetyt esimerkit viittaavat kohdassa 3.2.1 esitettyyn esimerkkiin.

Pyöristäminen (roll-up, drill-up, aggregation, consolidation) nostaa koostamisen tasoa. Esimerkiksi aika voitaisiin pyöristää vuosineljänneksiksi tai kokonaisiksi vuosiksi. Myyntialueet puolestaan voisi olla organisoitu osaksi valtioita, jolloin pyöristäminen yhdistäisi samaan valtioon kuuluvat alueet. Pyöristyssuhde voidaan määritellä myös eri ulottuvuuksien välille. Tuloksena OLAP-järjestelmä muodostaa uuden tietokuution. Tämän kuution ulottuvuudet vastaavat pyöristämisoperaation mukaisia määrityksiä ja sen arvot on laskettu uudestaan.

Porautuminen (drill down, roll down, drill through) laskee koostamisen tasoa. Tällöin käyttäjälle muodostetaan tietokuutio, jossa joko ulottuvuushierarkian tai ulottuvuuksien välille määritetyn suhteen mukaisesti näytetään atomisempaa tietoa. Esimerkiksi porautumalla syvemmälle ajan suhteen voitaisiin tarkastella myyntiä viikko- tai päivätasolla.

Viipaloinnissa (slice and dice, slicing) valitaan jostain ulottuvuudesta arvo, jonka perusteella muodostetaan osakuutio tarkasteltavaksi. Esimerkiksi valitsemalla arvo ”Alue 1” ulottuvuudesta myyntialueet, muodostuu kaksiulotteinen taulukkonäkymä, joka esittää alueella 1 tapahtuneet myynnit myyntialueittain ja kuukausittain.

Kiertämisen (pivot, rotate) avulla tietokuutio voidaan suunnata uudelleen. Visualisoinnin helpottamiseksi kuutio voidaan kääntää toisin päin, tai sen ulottuvuuksien paikkoja voidaan vaihtaa. Esimerkiksi aika- ja tuoteryhmäulottuvuuksien paikkoja voitaisiin vaihtaa, jos haluttaisiin tarkastella erityisesti ajan vaikutusta myyntiin. Ulottuvuuden tilalle voidaan myös vaihtaa uusi ulottuvuus. Esimerkiksi myyntialue voitaisiin korvata ulottuvuudella, joka määrittelee myynnin suorittaneen myyjän.

Valikoiminen (screening, selection, filtering) tarkoittaa tiedon tai ulottuvuuden jäsenten arviointia tietyn kriteerin mukaan. Kriteerin perusteella rajoitetaan mitä tietoa haetaan. Mukaan tuloksiin voidaan valita esimerkiksi vain 10 myydyintä tuoteryhmää.

4 TIEDON LOUHINTA

Tässä luvussa selitetään, mitä tiedon louhinta tarkoittaa sekä millaisia havaintoja ja tietoja sen avulla yritetään tuottaa. Yleisimmät tiedon louhinnassa käytetyt keinot, luokittelu, klusterointi ja assosiointi esitellään, jotta saadaan muodostettua kuva siitä, miten havaintoja pyritään saamaan aikaan.

4.1 Käsite ja tavoitteet

Tiedon louhinta on prosessi, jossa suurista tietokannoista erotetaan validia, ennestään tuntematonta ja kokonaisvaltaista toiminnan tueksi kelpaavaa tietoa, ja jossa tätä tietoa käytetään kriittisten liiketoimintapäätösten tekemiseen (Simoudis 1996).

Elmasrin ja Navathen (2007, 945) mukaan tiedon louhinta tarkoittaa uuden informaation louhintaa tai löytämistä, malleina tai sääntöinä suurista tietomääristä.

Näitä määritelmiä mukaillen tiedon louhinta määritellään tässä tutkielmassa tekniikaksi, jota käytetään löytämään uutta tietämystä suurista tietomääristä, ja tämän tietämyksen esittämistä kaavoina, malleina tai sääntöinä. Tätä tekniikkaa käytetään osana prosessia, jota kutsutaan termillä tietämyksen löytäminen tietokannasta.

Tietovarastoinnin tavoitteena on tarjota tietoa päätöksenteon tueksi. Tiedon louhintaa voidaan käyttää yhdessä tietovaraston kanssa. Tietoa voidaan louhia myös operatiivisista tietokannoista, joihin on tallennettu yksittäisiä tapahtumatietoja. Louhintaa voidaan kuitenkin tehostaa käyttämällä tietovarastoja, joissa tieto on koostettua ja yhteenvedettyä. Tiedon louhinta auttaa tunnistamaan uusia merkityksellisiä malleja, joita ei välttämättä pystytä löytämään pelkästään tekemällä kyselyjä tai prosessoimalla tietoa tai metatietoa tietovarastossa. (Elmasri & Navathe 2007, 946)

Elmasri ja Navathe (2007, 947) jakavat tiedon louhinnan tavoitteet ja käyttötarkoitukset karkeasti neljään luokkaan:

- Ennustaminen (prediction). Tiedon louhinnalla voidaan ennustaa, miten tietyt tiedon attribuuttien arvot tulevat käyttäytymään tulevaisuudessa. Tämän perusteella voidaan esimerkiksi ennustaa tulevia myyntimääriä, tai miten tietyt liiketoimintapäätökset tulisivat vaikuttamaan myyntiin. Myös esimerkiksi maanjäristyksiä voidaan ennustaa tutkimalla historiatietoa seismisestä toiminnasta.

- Tunnistaminen (identification). Kerätystä tiedosta löytyvien mallien ja kaavojen perusteella voidaan tunnistaa erilaisia kohteita, tapahtumia tai

toimintaa. Esimerkiksi tietojärjestelmään tunkeutujat voidaan tunnistaa heidän noudattamansa käyttäytymismallin mukaan. Tunnistamista voidaan hyödyntää myös mm. geenitutkimuksessa ja käyttäjien todentamisessa (authentication), joka on erityinen tunnistamisen muoto.

- Luokittelu (classification). Erilaisten parametrien perusteella voidaan tiedon louhinnan avulla löytää luokkia ja kategorioita. Luokittelua voidaan käyttää hyväksi yhdessä muiden tiedon louhinnan keinojen kanssa, joko luokittelemaan louhinnan tuloksia jälkeenpäin tai ennen muuta louhintaa jakamaan ongelmaa pienempiin osaongelmiin. Luokittelemalla voidaan esimerkiksi tunnistaa erilaisia asiakasryhmiä.

- Optimointi (optimization). Tiedon louhintaa voidaan myös hyödyntää resurssien, kuten ajan, tilan, rahan tai materiaalin, käytön optimointiin. Optimointia voidaan suorittaa esimerkiksi myynnin tai voiton maksimoimiseksi.

Kuten jo aiemmissa esimerkeissä on tullut esille, tiedon louhintaa voidaan käyttää useilla eri sovellusalueilla. Vähittäismyynti ja markkinointi, rahoitus ja pankkitoiminta, vakuutusala, tuotanto sekä terveydenhoito ovat näistä tärkeimpiä (Elmasri & Navathe 2007, 970 ja Connolly & Begg 2005, 1234).

4.2 Louhintatapoja

Tiedon louhinta on osa laajempaa prosessia, josta käytetään termiä tietämyksen löytäminen tietokannasta. Prosessi koostuu kuudesta vaiheesta, jotka ovat tiedon valinta (data selection), tiedon puhdistus (data cleansing), rikastus (enrichment), tiedon muuntaminen tai koodaus (data transformation, data encoding), tiedon louhiminen (data mining) sekä löydetyn informaation raportointi ja esittäminen (reporting and display). (Elmasri & Navathe 2007, 946) Jotta tietoa voidaan louhia, tulee tietoa siis esikäsitellä samaan tapaan kuin tietovarastoon varastoitavaa tietoa. Louhinnan tehostamisen lisäksi myös tämä puoltaa tietovaraston käyttöä tiedon louhinnan tukena.

Tiedon louhinta voidaan yhdistää myös OLAP-teknologioiden kanssa. Han (1998a) esittelee artikkelissaan käsitteen OLAP-louhinta (OLAP mining). Myöhemmässä artikkelissaan Han (1998b) käyttää OLAP:n ja tiedon louhinnan yhdistämisestä termiä OLAM (online analytical mining). OLAM mahdollistaa tiedon louhintaa vaativien havaintojen tuottamisen sekä niiden käsittelyn ja analysoinnin OLAP-operaatioiden avulla.

Eri sovellusalueilta on löydettävissä ongelmia, joihin voidaan etsiä ratkaisuja tiedon louhinnan avulla. Kutakin ongelmatyyppiä ratkaisemaan on määritetty algoritmeja, ja tiedon louhintaa luokitellaan usein näiden algoritmityyppien mukaan. Algoritmimäärityksistä kolme selvästi yleisintä ovat luokittelu

(classification), klusterointi (clustering) ja assosiaatiosäännöt (association rules). (Hellerstein & Stonebraker 2005, 650) Seuraavissa alakohdissa esitellään nämä tiedon louhinnan keinot pääpiirteittäin. Elmasri ja Navathe (2007, 949) mainitsevat edellä mainittujen kolmen tyypin lisäksi myös tiettyä järjestystä noudattavat kaavat (sequential patterns) ja kaavat aikasarjojen sisällä (patterns within time series). Heidän mukaansa tiedon louhinnalla tuotetut havainnot ovat useimmiten näiden viiden tyypin yhdistelmä.

4.2.1 Luokittelu

Luokittelu (classification) on tiedon louhinnan muoto, jossa tallennettu aineisto jaetaan ennalta määriteltyihin ja nimettyihin luokkiin. Luokittelua kutsutaan ohjatuksi oppimiseksi, sillä ennen luokittelua täytyy järjestelmään rakentaa opetusjoukko (training set), jonka perusteella loput aineistosta automaattisesti luokitellaan. Opetusjoukon jokainen tietue koostuu useasta kentästä tai atribuutista. Yksi attribuuteista on luokitteleva attribuutti, joka määrittelee mihin luokkaan opetusjoukon tietue kuuluu. (Shafer, Agrawal & Mehta 1996) Opetusjoukon luokittelun perusteella muodostetaan siis järjestelmälle malli siitä, miten muut attribuutit vaikuttavat luokittelevan attribuutin arvoon. Loput aineistosta sijoitetaan luokkiin tämän mallin mukaisesti.

Luokittelua voidaan käyttää hyväksi monilla eri tavoin. Esimerkiksi kirjeiden lajittelu, luottojen myöntäminen henkilöille taloudellisten ja henkilökohtaisten tietojen perusteella ja alustavien diagnoosien tekeminen potilaan oireiden perusteella voidaan suorittaa luokittelun avulla (Michie, Spiegelhalter & Taylor 1994).

4.2.2 Klusterointi

Klusterointi (clustering) on tietokannan tapahtumien ohjaamatonta luokittelua ryhmiksi (group, cluster) (Jain, Murty & Flynn 1999). Tutkittava aineisto voidaan nähdä moniulotteisena joukkona pisteitä, missä pisteet eivät ole jakaantuneet tasaisesti. Klusteroinnilla voidaan tunnistaa tästä pisteavaruudesta harvat ja tiheät kohdat ja muodostaa niistä ryhmiä. (Zhang, Ramakrishnan & Livny 1996) Klusteroinnissa ei siis muodosteta opetusjoukkoa kuten luokittelussa, vaan ryhmittely tehdään laskemalla tietokannan tapahtumien yhtäläisyyksiä. Tämä voidaan määrittää esimerkiksi laskemalla tapahtumien Euklidinen etäisyys niiden klusterointiin valittujen kenttien arvojen perusteella (Elmasri & Navathe 2007, 964).

Jain ym. (1999) mainitsevat segmentoinnin, ennustavan mallintamisen ja visualisoinnin eräinä tiedon louhinnan lähestymistapoina, jotka hyödyntävät klusterointia. Seuraavassa annetaan lyhyet kuvaukset näistä heidän mukaansa.

Klusterointimenetelmiä käytetään tiedon louhinnassa tietokantojen segmentoimiseen homogeenisiksi ryhmiksi. Tätä voidaan hyödyntää tiedon pakkaamisessa, koska se mahdollistaa työskentelyn klusterien parissa yksittäisten tietueiden sijasta. Klusteroinnin avulla voidaan myös tunnistaa alijoukkojen ominaisuuksia, ja käyttää näitä tietoja esimerkiksi markkinointiin joka on suunnattu tietylle asiakasryhmälle.

Tiedon louhinta voi auttaa tilastollisen analyysin tekijää löytämään mahdollisia hypoteeseja ennen tilastollisten työkalujen käyttämistä. Ennustava mallintaminen käyttää klusterointia ryhmittelemään aineistoa, päättelee sääntöjä sekä esittää malleja kuvaamaan näitä ryhmiä. Esimerkiksi lehden tilaajat voitaisiin ensin klusteroida ja tämän jälkeen päätellä automaattisesti säännöt, joiden perusteella tilaajat eroteltaisiin tilauksensa todennäköisesti uusiviin ja lopettaviin (Simoudis 1996).

Klustereita isoissa tietokannoissa voidaan käyttää tiedon visualisointiin. Näin voidaan auttaa analyytikoita tunnistamaan ryhmiä ja niiden aliryhmiä, joilla on samanlaisia ominaisuuksia.

4.2.3 Assosiaatiosäännöt

Assosiaatiosäännöt ovat tallennetusta tiedosta löytyviä assosiaatioita tapahtumien välillä. Assosiaatiosääntö (association rule) on muotoa X => Y, millä tarkoitetaan sitä, että on olemassa yhteyksiä tapahtumien X ja Y välillä. Esimerkiksi asiakas, joka ostaa tuotteen X, ostaa usein myös tuotteen Y. Koska tällaisia yhteyksiä voi olla lukemattomia, kiinnostavat assosiaatiosäännöt erotellaan muista käyttämällä kahta mittaria, tuki (support, prevalence) ja luottamus (confidence, strength).

Assosiaatiosäännön X => Y tuki ilmaisee sen, miten usein tietty sääntö esiintyy aineistossa. Se ilmoittaa, kuinka monessa prosentissa tutkittavan aineiston tapahtumista sääntö löytyy. Jos esimerkiksi kolme kymmenestä asiakkaasta on ostanut sekä maitoa että juustoa, on assosiaatiosäännön maito => juusto tuki 30 %.

Säännön X => Y luottamus ilmaisee, kuinka moni tapauksista X johtaa tapahtumaan Y. Jos edellä mainitussa esimerkissä maitoa ostaneita asiakkaita oli yhteensä 6, ja heistä 3 osti myös osti myös juustoa, on säännön maito => juusto luottamus 50 %.

Seuraavassa kuvataan erilaisia assosiaatiosääntötyyppejä Elmasrin ja Navathen (2007) mukaan.

Ostoskorianalyysi

Ostoskorianalyysi käsittelee ongelmaa, jossa pyritään suuresta määrästä ostoskorityyppistä tietoa löytämään assosiaatiosääntöjä, joilla on tietty minimiluottamus ja minimituki. Tällaisia ongelmia ovat mm. tavaratalon asiakkaiden tekemien ostosten analysointi tai vakuutusyhtiön asiakkaiden ostamien vakuutusten välisten yhteyksien tutkiminen. Tämän tyyppisillä assosiaatiosäännöillä voidaan siis tuottaa lausuntoja, kuten ”90 % niistä, jotka ostavat leipää ja voita, ostavat myös maitoa”. (Agrawal, Imieliński & Swami, 1993) Löydettyjen yhteyksien perusteella voidaan esimerkiksi suunnitella alennusmyyntejä ja mainontaa, tai parantaa tuotteiden asettelua siten, että se tukee asiakkaiden ostokäyttäytymistä.

Assosiaatiosäännöt hierarkioiden välillä

Usein on mahdollista ryhmitellä tuotteet ja palvelut hierarkioiksi. Esimerkiksi ruokakaupan tuotteet voidaan jakaa ensin korkeammalla tasolla juomiin ja jälkiruokiin, ja nämä taas edelleen alemmilla tasoilla vaikkapa hiilihapollisiin ja – hapottomiin sekä jäätelöihin ja leivoksiin. Vaikka ylemmän tason assosiaatiosääntöjä juomat => jälkiruoat tai jälkiruoat => juomat, ei löytyisikään, voivat alemmat hierarkiatasot muodostaa tueltaan ja luottamukseltaan riittävän vahvoja sääntöjä. Jos sovellusalueella on luonnollinen hierarkkinen luokitus, sen sisältä löydetyt assosiaatiot eivät ole erityisen kiinnostavia. Sen sijaan ovat hierarkioiden välillä esiintyvät assosiaatiot voivat olla kiinnostasvia. (Elmasri & Navathe 2007, 957-958)

Negatiiviset assosiaatiot

Negatiiviset assosiaatiot ovat tyyppiä X => ¬Y tai ¬X => Y. Esimerkiksi sääntö maito => ¬juusto, jolla on 50 % luottamus, voitaisiin tulkita seuraavasti: ”50 % asiakkaista, jotka ostavat maitoa, eivät osta juustoa”.

Jos tietokannassa on 10000 tapahtumatietoa, on niille olemassa 2^{10000} mahdollista yhdistelmää, joista suurin osa ei esiinny kertaakaan tietokannassa. Negatiivisten assosiaatiosääntöjen ongelma on siis se, että kiinnostavia sääntöjä on vaikea löytää kaikkien mahdollisten assosiaatioiden joukosta. (Elmasri & Navathe 2007, 959)

Moniulotteiset assosiaatiot

Yllä kuvattuja assosiaatioita sanotaan yksiulotteisiksi assosiaatioiksi. Ne käsittelevät assosiaatioita kuten ostettu_tuote(maito) => ostettu_tuote(juusto), jossa ulottuvuus siis käsittelee ostettuja tuotteita. Elmasrin ja Navathen (2007, 958) esimerkkinä moniulotteisesta assosiaatiosta on aika(6.30 … 8.00) => ostettu_tuote(maito), jossa tutkitaan assosiaatioita ostotapahtuman ajan ja ostetun tuotteen välillä.

Lu, Feng ja Han (2000) laajentavat artikkelissaan moniulotteisia assosiaatioita myös erillisten tapahtumien välille. He esittävät, että perinteisten yksiulotteisten tapahtumien sisäisten assosiaatioiden (intratransaction associations) lisäksi voidaan löytää monimutkaisempia assosiaatioita, kuten ”Kahden kuukauden sisällä siitä, kun McDonald ja Burger King avaavat haaraliikkeen, KFC avaa todennäköisesti haaraliikkeen alle mailin päähän.”

5 YHTEENVETO

Tässä tutkielmassa tarkasteltiin kolmea päätöksentekoon liittyvää aihealuetta, tietovarastointia, OLAP:ia ja tiedon louhintaa. Kustakin aihealueesta pyrittiin antamaan yleiskuvaus ja selvittämään yhteyksiä niiden välillä. Erityistä huomiota kiinnitettiin käsitteiden yhdenmukaiseen määrittämiseen.

Tietovarastointi on prosessi, jossa organisaation toiminnan yhteydessä syntyvää tietoa yhdistetään useista operatiivisista tietolähteistä tietovarastoon asiakassovellusten käytettäväksi. Nämä suuret tietomassat ovat tehokkaammin hyödynnettävissä tietovaraston kautta, sen sijaan että analyytikot ja päätöksentekijät tekisivät kyselyjä suoraan operatiivisiin tietokantoihin.

OLAP määriteltiin tässä tutkielmassa lähestymistavaksi, jonka mukaisesti toteutettua järjestelmää voidaan käyttää hyväksi organisaation tietovarantojen analysoinnissa. OLAP–järjestelmä sisältää yleensä OLAP-palvelimen, johon tieto tallennetaan käyttämällä moniulotteista tietomallia. Asiakassovellukset tekevät kyselyitä tälle palvelimelle ja visualisoivat tietoa yleensä tietokuutioina. Tulosten käsittelyyn ja analysointiin käytetään OLAP-operaatioita. Tietovarastoinnilla ja OLAP:lla on useita yhteisiä piirteitä, ja kirjallisuus käsitteleekin näitä usein yhdessä. Tietovarastoissa käytetään sekä perinteisiä relaatiotietokantapalvelimia että erilaisia moniulotteisesti tietoa tallentavia palvelimia, jotka on kehitetty pääasiassa OLAP:n tarpeita silmällä pitäen. OLAP-järjestelmä voikin käyttää tiedon tallentamiseen ja käsittelyyn tietovarastoa, tai yksittäistä OLAP-palvelinta. Tietovarastoarkkitehtuurin osana OLAP-palvelin toimii yleensä paikallisvarastona.

Tiedon louhinta on osa prosessia, jota kutsutaan termillä tietämyksen löytäminen tietokannasta. Tiedon louhinta on tekniikka, jolla pyritään paikantamaan ja johtamaan tallennetusta tiedosta uutta tietämystä, joka voidaan esittää esimerkiksi kaavoina, malleina tai sääntöinä. Tiedon louhintaa voidaan soveltaa suoraan operatiivisiin tietokantoihin, mutta tällöin joudutaan suorittamaan samat tiedon esikäsittelyn vaiheet kuin tietovarastoinnissa. Usein onkin siis järkevämpää louhia tietoa organisaation tietovarastosta. Tiedon louhinta voi olla myös tehokkaampaa, jos varastoitu tieto on koostettu louhintaa tukevalla tavalla. Tiedon louhinta voidaan yhdistää myös OLAP-teknologioiden kanssa. Tästä käytetään termiä OLAP-louhinta tai OLAM.

Tietovarastointi, OLAP ja tiedon louhinta ovat ratkaisuja organisaatioiden tarpeisiin hyödyntää yhä lisääntyvää operatiivista liiketoimintatietoa päätöksenteon tukena. Vaikka näistä on käsitteistä olemassa kirjallisuutta, on niiden käyttämä termistö monilta osin epäyhtenäistä. Tämä johtunee siitä, että varsinkin tietovarastointi ja OLAP ovat kehittyneet ensin kaupallisten yritysten toimesta ja tieteellinen tutkimus niistä on nuorempaa. Tässä tutkielmassa käsitteitä pyrittiin yhdenmukaistamaan muodostamalla niistä yleiskuvaus.

Jatkotutkimuksissa kutakin näistä käsitteistä voisi tarkastella tarkemmin ja selkeyttää aihepiirejä entisestään. Myös tietovaraston suunnittelu sekä tietokannan hallintajärjestelmien tarjoama tuki OLAP:lle ja tiedon louhinnalle ovat mielenkiintoisia aiheita jatkotutkimuksille.

LÄHDELUETTELO

Agrawal, R., Inmieliński, T., & Swami, A. 1993. Mining association rules between sets of items in large databases. Teoksessa P. Buneman and S. Jajodia (toim.) Proceedings of the 1993 ACM SIGMOD international Conference on Management of Data, Washington, D.C., United States, May 25-28, 1993. New York (NY): ACM Press, 207-216.

Barquin, R. 1996. On the First Issue of The Journal of Data Warehousing. The Journal of Data Warehousing 1, 2-6.

Chaudhuri, S. & Dayal, U. 1997. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26(1), 65-74.

Codd, E.F., Codd, S.B., Salley, C.T. 1993. Providing OLAP to User-Analysts: An IT Mandate. Codd E.F. & Associates. Technical report.

Connolly, T.M. & Begg, C.E. 2005. Database Systems – A Practical Approach to Design, Implementation, and Management. Harlow, England: Addison Wesley.

Elmasri, R. & Navathe, S.B. 2007. Fundamentals of database systems. Boston (MA): Addison-Wesley.

Han, J. 1998a. OLAP Mining: An Integration of OLAP with Data Mining. Teoksessa S. Spaccapietra & F. Maryanski (Toim.) Data Mining and Reverse Engineering: Searching for Semantics, IFIP TC2/WG2.6 Seventh Conference on Database Semantics (DS-7), October 7-10, 1997, Leysin, Switzerland. Chapman & Hall, 3-20.

Han, J. 1998b. Towards on-line analytical mining in large databases. ACM SIGMOD Record. 27(1), 97-107.

Hellerstein, J.M. & Stonebraker, M. (toim.) 2005. Readings in Database Systems. Cambridge (MA): MIT Press.

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

Inmon, W.H., Welch, J.D. & Glassey, K.L. 1997. Managing the Data Warehouse. New York: Wiley.

Jain, A. K., Murty, M. N., & Flynn, P. J. 1999. Data clustering: a review. ACM Computing Surveys 31(3), 264-323.

Jarke, M. , Lenzerini, M. , Vassiliou, Y. & Vassiliadis, P. 2000. Fundamentals of Data Warehouses. Berlin: Springer.

Lu, H., Feng, L., & Han, J. 2000. Beyond intratransaction association analysis: mining multidimensional intertransaction association rules. ACM Transactions on Information Systems 18(4), 423-454.

OLAP Council 1997. OLAP Council White Paper [online]. OLAP Council [viitattu 10.9.2007]. Saatavilla www-osoitteessa <http://www.olapcouncil.org/research/whtpaply.htm>.

McFadden, F.R. , Hoffer, J.A. & Prescott, M.B. 1999. Modern database management. Reading (MA): Addison Wesley Longman.

Michie, D., Spiegelhalter, D.J. & Taylor, C.C. 1994. Introduction. Teoksessa D. Michie, D.J. Spiegelhalter, C.C. Taylor & J. Cambell (toim) 1994. Machine Learning, Neural and Statistical Classification. Upper Saddle River (NJ): Ellis Horwood.

Shafer, J.C., Agrawal, R. & Mehta, M. 1996. SPRINT: A Scalable Parallel Classifier for Data Mining. Teoksessa T. M. Vijayaraman, A.P. Buchmann, C. Mohan & N.L. Sarda (toim.) Proceedings of the 22th International Conference on Very Large Data Bases, September 3-6, 1996. San Francisco (CA): Morgan Kaufmann Publishers Inc, 544-555.

Simoudis, E. 1996. Reality Check for Data Mining. IEEE Expert: Intelligent Systems and Their Applications 11(5), 26-33.

Zhang, T., Ramakrishnan, R. & Livny, M. 1996. BIRCH: an efficient data clustering method for very large databases. Teoksessa J. Widom (toim.) Proceedings of the 1996 ACM SIGMOD international conference on Management of data, Montreal, Quebec, Canada, June 4-6, 1996. New York(NY): ACM Press.