Teksti: John Martin

FM John Martin, Ineo Corporate Performance Management Oy, Turku

Kirjoittaja toimii konsulttina Ineo Groupin Corporate Performance Management yksikössä.

Uusi tietovarastojen mallinnusmenetelmä – Data Vault

Data Vault-mallin syntyhistoria

Alussa oli Codd & Date, jotka kehittivät ns. relaatiomallin (3NF) OLTP-järjestelmiä varten. 80-luvun alussa Inmon muokkasi tätä relaatiomallia, jotta se paremmin sopisi tiedon varastointiin. Käytännössä malliin lisättiin aikaleima avainkenttiin. Muutamia vuosia myöhemmin Kimball kehitti tähtimallin. Tähtimallin etuja ovat mm. helppokäyttöisyys, aggregaatiot, nopeat kyselyt, tuki OLTP-järjestelmille ja mahdollisuus muuttaa tietomallia. Tätä yhden aihealueen tähtimallia kutsutaan paikallisvarastoksi (data mart). Lisääntyvän tietovarastointitarpeen tyydyttämiseksi näitä yhden aihealueen tähtimalleja alettiin yhdistellä suuremmiksi useamman aihealueen kokonaisuuksiksi (conformed data marts).

Myöhemmin huomattiin että relaatiomallilla ja tähtimallilla esiintyi suorituskykyongelmia ja muita heikkouksia kun datamäärät lisääntyivät. Tämän ongelman ratkaisemiseksi Dan Linstedt aloitti uuden nimenomaan tietovarastointiin tarkoitetun mallin kehittämisen, jonka hän sai valmiiksi vuosituhannen vaihteessa. Mallia kutsutaan nimellä Data Vault.

Miksi Data Vault?

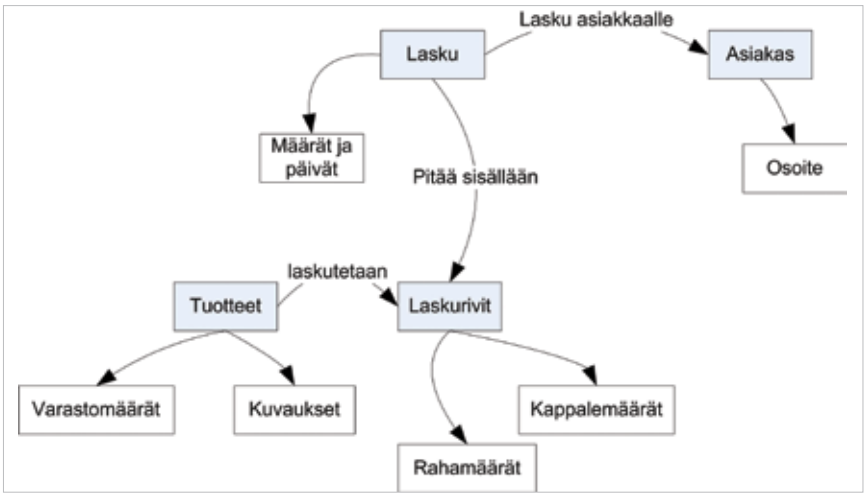
Data Vault-malli mukautuu nopeasti yrityksen liiketoiminnan muutoksiin mahdollistaen liiketoiminnan ja IT:n yhteistyön yhteisten tavoitteiden saavuttamiseksi. Data Vaultin-mallin avulla voidaan löytää mahdollisia liiketoiminnan ongelmakohtia, joita aiemmin ei ole huomattu. Muutoshallinta helpottuu huomattavasti verrattuna perinteisiin tietovarastointimenetelmiin. Uusien liiketoimintayksiköiden lisääminen malliin on helppoa. Erilaisten datalähteiden lisääminen on mahdollista. Kaikki tietovarastoon tullut data voidaan jäljittää lähdejärjestelmään.

Data Vault-malli perustuu Inmonin ja Kimballin malleihin ja on eräänlainen parhaista paloista koottu hybridi em. malleista. Ensimmäinen ja ainoa varta vasten tietovarastointiin kehitetty malli.

Koska DV-malli perustuu vahvasti liiketominnan prosesseihin, kannattaa mallinnustyö aloittaa kuvaamalla itse liiketoimintaprosessi.

Laskutus on yritykselle liiketoiminnallisesti tärkeä prosessi, josta yksinkertaistettu kuvaus alla olevassa malliesimerkissä:

Asiakas on tilannut tuotteita ja tämä tilaus laskutetaan. Lasku lähetetään asiakkaalle. Laskulla on laskurivejä (kappalemääriä, rahasummia yms.). Tuotteilla on varastosaldoja, kuvauksia yms.



Kuva 1. Liiketoiminta-prosessin kuvaaminen.

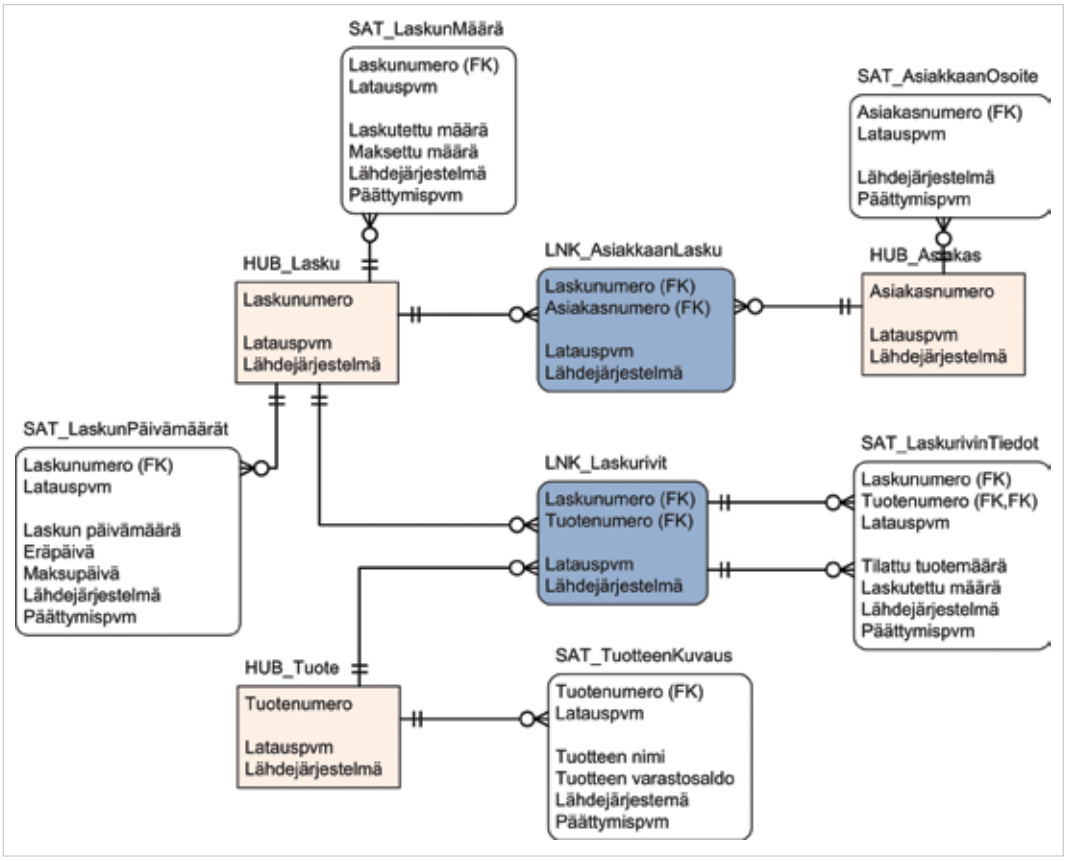
DV:n palaset

Jotta arkkitehtuuri olisi yksinkertainen ja tyylikäs DV-malli koostuu kolmesta pääentiteetistä; Hubi, Linkki ja Satelliitti (Hub, Link ja Satellite). Hubeilla kuvataan yritykselle tärkeitä liiketoiminta-avaimia (vrt. dimensiot). Linkit kuvaavat hubien välisiä relaatioita, rooleja yms. ja satelliitit antavat liiketoiminta-avaimille kontekstin (vrt. fakta).

Hub-entiteetit

Hub-taulu on entiteetti joka pitää sisällään uniikin listan yrityksen yhdestä liiketoiminta-avaimesta (business key). Liiketoiminta-avaimia ovat esim. laskunumero, työntekijän numero, tuotenumero, toimipaikka jne. Jos yritys kadottaisi liiketoiminta-avaimen katoasi myös yhteys liiketoiminnan kontekstiin tai ympäröivään informaatioon. Muita hubin attribuutteja ovat mm. surrogaatti-avain, datan lähde, latauspäivämäärä (vrt. esim ajonumero)

Data Vaultia rakennettaessa aloitetaan aina hubien määrittelyllä. Kun hubit ovat saatu määriteltyä voidaan aloittaa linkkien määrittely.



Kuva 2. Kun liiketoiminta-prosessi (kuva 1) kuvataan Data Vault-mallilla saadaan seuraava tietokantamalli.

Link-entiteetit

Linkit ovat yhdistäviä tauluja, jotka yksilöivät yhteydet liiketoiminta-avaimien välillä. Linkeillä kuvataan transaktioita, rooleja ja muita tapahtumia. Linkkien avulla määritellään myös tietovaraston karkeisuutta (grain). Linkeillä voidaan myös tarvittaessa kuvata hierarkioita. Viiteavaimina käytetään hubien ja/tai linkkien perusavaimia. Muita linkin attribuutteja ovat latauspäivämäärä ja datan lähde.

1. **Data Vaultin rakentaminen**
   Mallinna hubit. Hubien mallintaminen vaatii ymmärryksen liiketoiminnasta puitealueella (scope)
2. Muodosta linkit. Yhdistä liiketoiminta-avaimet toisiinsa eli muodosta näkemys siitä miten liiketoiminta-avaimet ovat yhteydessä toisiinsa
3. Mallinna satelliitit. Anna liiketoiminta-avaimille ja hubieja yhdistäville transaktioille (linkit) konteksti.
4. Nyt alkaa koko liiketoiminta hahmottua.
   Mallinna mahdolliset Stand-alone-taulut (referenssitaulut), kuten kalenteri/aika koodi/kuvaus taulut. Eli tauluja joilla ei haluta ylläpitää historiaa mutta kuitenkin käytetään ja tarvitaan. Stand-alone-tauluja voivat olla myös normalisoidut kooditaulut tms, jo-
5. den avulla helpotetaan kyselyitä paikallisvarastoissa. Tarvittaessa optimoi suorituskykyä lisäämällä ns. Bridge-tauluja ja Point In Time-rakenteita

Satelliitti-entiteetti

Satelliiteihin tallennetaan liiketoiminta-avaimia kuvaavaa historioitavaa dataa. Kaikella satellitin datalla on taipumus muuttua ajan mukaan. Tästä johtuen rakeenteen tulee olla joustava ja pystyä tallentamaan kaikki muutokset toivotulla karkeisuudella. Esimerkiksi tuotteen nimi voi vaihtua (tuotenumeron pysyessä samana) ja tämä pitää saada kuvattua. Saman käsitteen satelliitit jaetaan muuttumistiheyden mukaan. Satelliitin perusavain muodostetaan hubin tai linkin viiteavaimesta ja latauspäivämäärästä. Muita satelliitin komponentteja ovat datan lähde-kenttä (record source).

Muutosjoustavuus

Edellä mallinettu yritys päättääkin luopua tuotteistaan kokonaan ja aloittaa palveluiden tuottamisen. Luodaan uusi hubi palveluille ja uusi linkki, jolla yhdistetään palvelut laskuihin ja päätetään tuotteiden päivittäminen lisäämällä päättymispäivämäärät (Load End Dates). Tuotteisiin liittyvät taulut jätetään luonnollisesti tietovaraastoon.

Yrityksen asiakkastietohin lisätään attribuutti. Tästä johtuen malliin lisätään esim. uusi satelliitti asikashubiin tai päivitetään olemassa olevaa satelliittia.

Helppoa.

Historiointi rakenteessa mukana.

Mikäli tietovarasto toteutetaan Data Vault mallina ei datan historioinnista tarvitse erikseen huolehtia. Historiointi on ”sisäänrakennettuna”. Pidetään vain huolta, että em. päättymispäivämääriä käytetään.

Skaalautuvuus

Pienin mahdollinen Data vault on 1 hubi ja siihen satelliitti ja suurin on laitteistosta riippuvainen (tera jopa petatavun kokoiset).

Muuta

Data Vault-mallilla toteutettu tietovarasto ei sellaisenaan sovellu käyttäjen kyselyalustaksi. Rakenne on normalisoitu mistä johtuu että kyselyt muodostuvat monimutkaisiksi. Edellyttää paikallisvarastoja. Sen sijaan Data Vault-tietovarasto sopii mainioisti tiedon luohinnan alustaksi (Data Mining).

Valitettavasti Data Vault-mallista ei vielä ole kirjaa saatavilla, joten parhaaksi Data Vault-lähteeksi jää www.danlinsted.com.