IIO30100 Tietokantojen suunnittelu

Tietovarastojen suunnittelu

kirjan Hovi, Huotari, Lahdenmäki: Tietokantojen suunnittelu & indeksointi, Docendo (2003, 2005) luku 8

© Jouni Huotari & Ari Hovi

Määrittely- ja pohdiskelutehtävä

- Mikä on tietovarasto?
- Miten termi Business Intelligence liittyy tietovarastointiin?
- Miten käsite data mart liittyy tietovarasttoihin (data warehouse)? Entä OLAP (online analytical processing)?
- Mikä on ETL-tekniikan perusidea? Mistä sanoista ETL tulee?

Tietovarastokannan suunnitteluperiaatteita

Tietovarastotietokannan tavoitteena on tukea tietojen kyselyä, analysointia ja raportointia

- Tämä on operatiivisissa kannoissa usein hankalaa (mutta ratkaisujakin on; osaatko mainita mitä?)

=> on tärkeää, että:

- historia säilytetään: käyttäjät haluavat tehdä trendianalyysiä
- eri operatiivisista järjestelmistä tulevat tiedot yhdenmukaistaan: yhdenmukaiset koodit ja perusavaimet
- tietorakenne on selkeä ja intuitiivinen: ymmärrettävät taulujen ja sarakkeiden nimet, kyselyjä on helppo tehdä

Tietovarastokannan suunnitteluperiaatteita

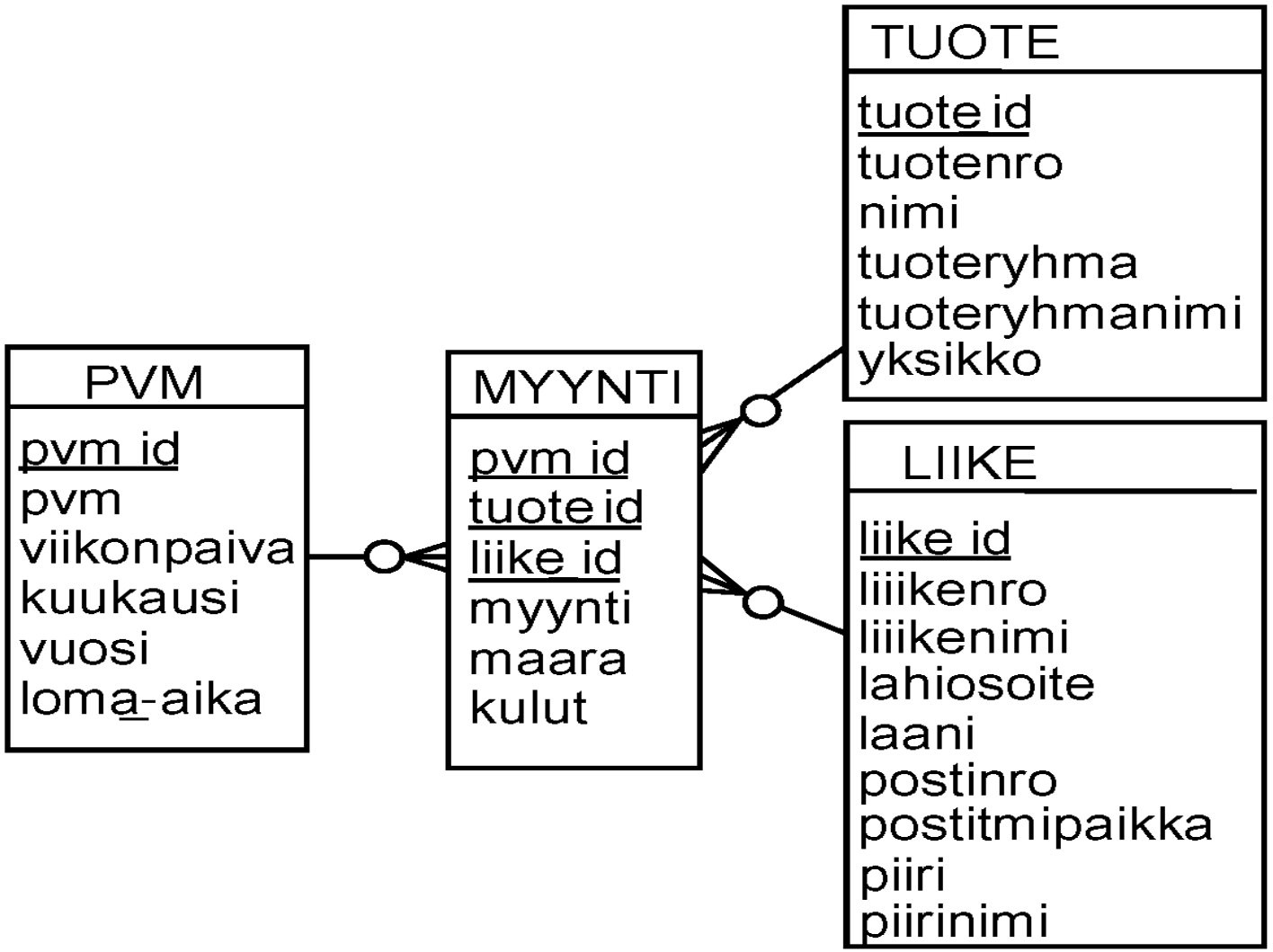
Ratkaisuja

- Tiedon "jalostus": tietoa on summattu valmiiksi, yleisiä tunnuslukuja on laskettu valmiiksi
- Tiedot on tallennettu riittävän tarkalla tasolla (esim. viikkotaso)
- denormalisointi ja summaus => saadaan riittävän hyvä vastausaika kyselyille (tähtimalli)

Tähtimalli

- Tähtimalli (star schema, dimensional modelling) eli dimensiomalli on tietovarastojen yhteydessä esitetty uusi tapa soveltaa käsiteanalyysiä
- Tähtimallissa on luovuttu osittain kolmannesta normaalimuodosta kyselyjen yksinkertaistamiseksi ja nopeuttamiseksi (denormalisointi)
- Hintana tästä on levytilan tietty lisätarve (yleensä melko vähäinen) sekä se, että tietovaraston lataaminen monimutkaistuu jonkin verran
- Tähtimalli sopii hyvin erityisesti numeerisen ja moniulotteisen tietoaineksen mallintamiseen (esim. päiväys voidaan jakaa monelle tasolle => aikaulottuvuus)
- Tähtimallia käytetään etenkin pienempien tietovarastokantojen, ns. paikallisvarastojen (data mart) suunnittelussa.
- Nimi tähtimalli tulee siitä, että rakenne muistuttaa tähteä sakaroineen (ks. kirjan kuva 8.1); jos sakarat jakautuvat edelleen pienempiin osiin, puhutaan lumihiutalemallista (kuva 8.2)
- Tähtimallin avulla suunnitellaan myös ns. OLAP-kantoja (Online Analytical Processing) ja ns. kuutioita (cube), jotka eivät ole tavanomaisia relaatiokantoja vaan tietojen analysointiin tarkoitettuja erikoiskantoja

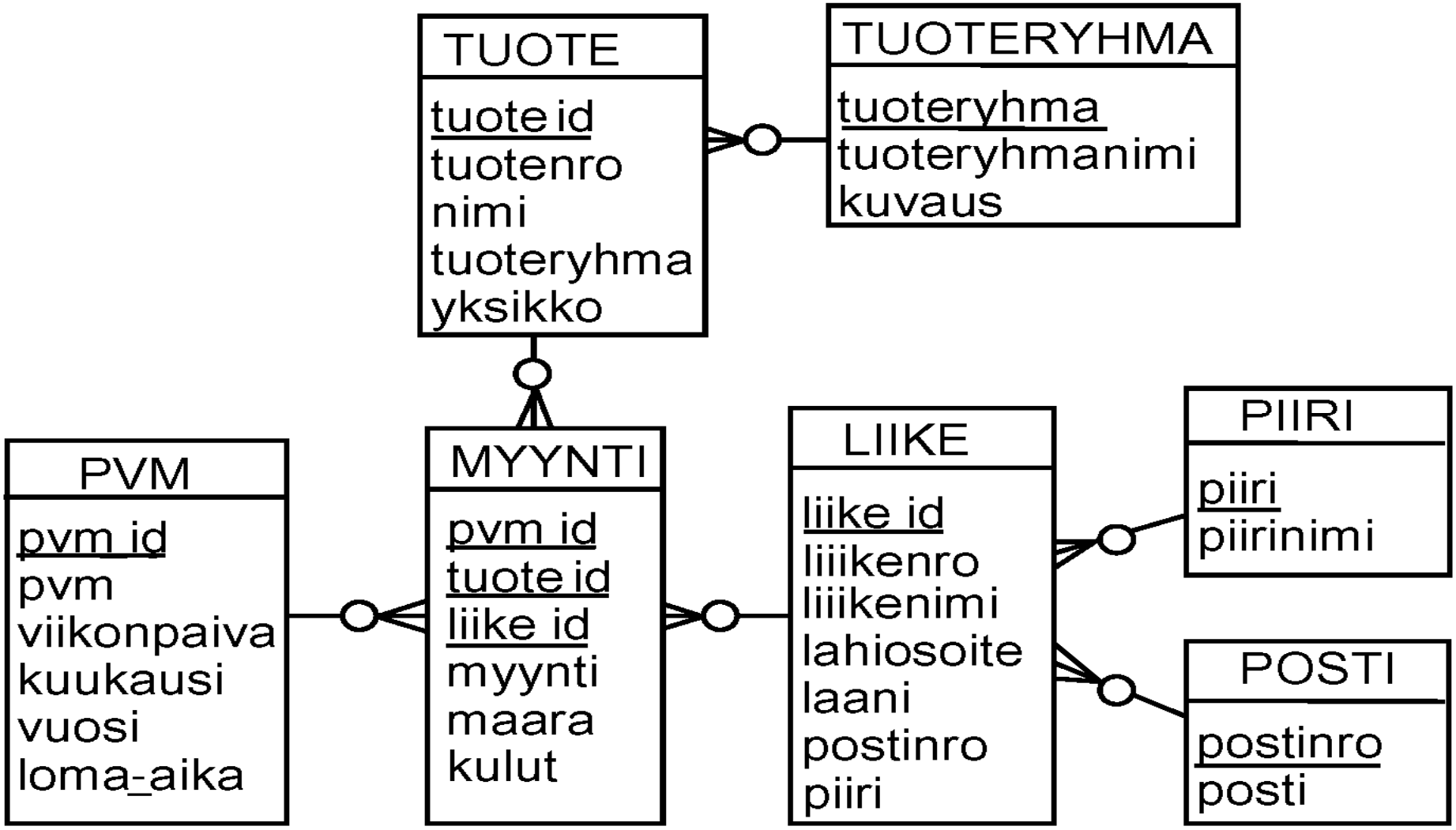
Kuva 8.1: Tähtimalli



Tähtimalli: faktataulu ja dimensiotaulut

- Tähtimallissa on keskellä iso taulu, ns. faktataulu, ja ympärillä useita pienempiä dimensiotauluja eli ulottuvuustauluja
- Faktataulun tiedot ovat luonteeltaan tapahtumatyyppisiä, kuten myyntitapahtumia
- Numeroarvoja (esimerkiksi myynti, kpl, vero, saldo) sisältäviä sarakkeita kutsutaan faktoiksi
- Perusavain koostuu dimensiotaulujen perusavaimista yhdisteltynä
- Faktataulun numeroarvoja lasketaan yleensä yhteen erilaisilla dimensioista saaduilla tasoilla; voidaan myös laskea keskiarvoja ja lukumääriä eri dimensioilla
- Faktataulu on karkeisuudesta riippuen yleensä iso, esim. miljoonia rivejä
- Tähtimallikyselyä tehtäessä dimensiotaulut yhdistetään faktatauluun liitoksilla
- Dimensiotauluissa säilytetään tapahtumiin verrattuna staattisia tietoja, joista saadaan erilaisia ryhmittelyjä ja hakukenttiä faktataulun rivejä summaaville kyselyille
- Tyypillisiä faktatauluja ovat asiakas, tuote, pvm, organisaatioyksikkö,
- Dimensiotaulut ovat faktatauluun verrattuna pieniä, usein sadoista tuhansiin riveihin.

Kuva 8.2: Lumihiutalemalli



Loput esimerkit, ks. kirja ja linkit

- http://www.dmreview.com/channels/dw_basics.html
- http://en.wikipedia.org/wiki/Data_warehouse
- http://en.wikipedia.org/wiki/Data_mart
- http://en.wikipedia.org/wiki/Star_schema
- http://en.wikipedia.org/wiki/Snowflake_schema
- http://en.wikipedia.org/wiki/OLAP