

Data Warehouse

Duomenų saugykla tai į dalyką orientuotas (dalykinis), integruotas, nekintantis ir skirtingas laike duomenų rinkinys, naudojamas vadybininkų sprendimams paremti (*angl. A data warehouse is a subject-oriented, integrated, nonvolatile, and time-variant collection of data in support of management's decisions*). Duomenų saugyklose duomenys yra apibendrinti (*angl. granular corporate data*). Šitie duomenys gali būti naudojami įvairiems tikslams.

Duomenų dalykinė sritis – duomenys yra suburti aplink tam tikrą įmonės dalyką (pvz., klientą, produktą, pardavimą ...)

Typical Subjects (tipiniai dalykai)

Commercial (*komercija*) {Customer (*klientas*), Product (*produktas*), Organizational unit (*organizacinis vienetas*), Employee (*darbuotojas*), Sales (*pardavimai*)}

Government Services (*valstybės tarnyba*) {Citizen (*piliėtis*), Department (*skyrius*), Employee (*darbuotojas*), Asset (*turtas*), Service (*tarnyba*)}

Academic (*universitetas*) {Student (*studentas*), Instructor (*vadovas*), Curriculum (*mokymo programa*), Degree (*mokslinis laipsnis*), Department (*katedra/skyrius*)}

Retail Bank (*bankas*) {Customer (*klientas*), Application (*pareiškimas*), Product (*produktas*), Account (*sąskaita*), Transaction (*transakcija*), Employee (*darbuotojas*), Organizational unit (*organizacinis vienetas*)}

- Identifikuoti pagrindines dalykines sritis (tam galima naudoti ER modelį, sukurtą pirmame lab. darbe).
- Aprašyti kiekvieną dalyką (*angl. subject*) lentelėje:

Pavadinimas	Aprašymas	Ryšys su kitais dalykais (trumpas aprašymas)	Dalyko tipas (dimensija, faktas)	Kitos pastabos

- Pateikti apibendrintą dalykinį modelį (geras yra antras paveikslėlis).

Žemiau yra pateikti keli modelių pavyzdžiai.

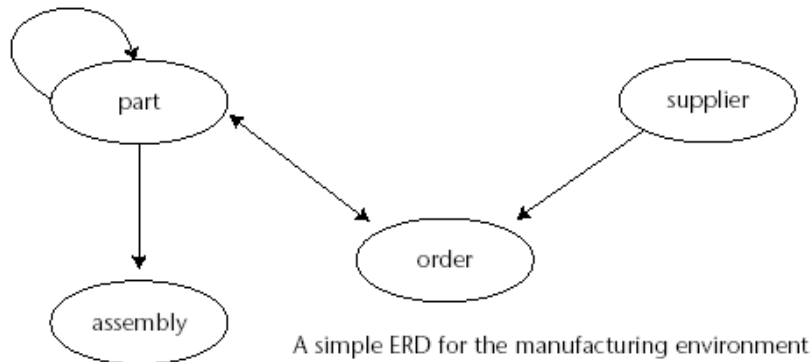
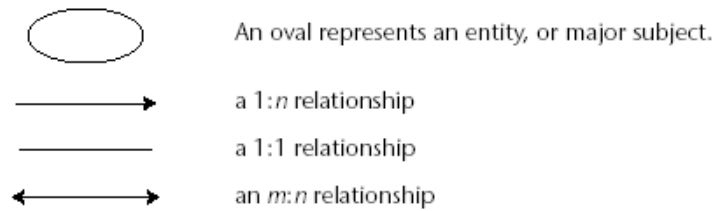
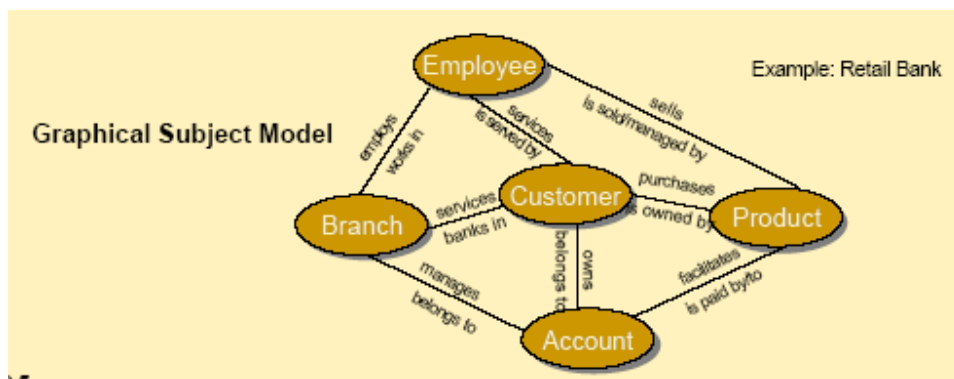


Figure 3-10 Representing entities and relationships.



Pagrindinės aukščiausio lygmens modelio savybės (*angl. features*) yra esybės ir ryšiai.

Corporate Subject Model Tips

- All subjects must be related to at least one other subject
 - If not, you have missed either a subject or a relationship!
- Subject definitions should be consistent across business units
 - They often provide dimensions in later, more detailed models
 - Dimension conformity is important for the detailed models
- A good model is easily understood
 - Fits on one page
 - Has between 10 and 20 subjects (varies by complexity of organisation)
 - Is described in business terms
 - Can be clearly presented in graphical format

A relationship - Associates two entities

- Includes a relationship label describing the relationships from both directions

- Indicates cardinality - the number of occurrences of the entities that relate e.g.
 - One-to-one
 - One-to-many
 - Many-to-many
- Indicates optionality - whether an occurrence of the related entity is:
 - Mandatory (at least one)
 - Optional (can be zero).
- Can be defined precisely using formal notation.

Duomenų integravimas - Kita svarbi duomenų saugyklos savybė – duomenys duomenų saugykloje yra integruoti. Integracija yra labai svarbus aspektas duomenų saugyklose, kadangi duomenys yra imami iš daugelio šaltinių. Duomenys yra konvertuojami, formatuojami, išdėstomi nuosekliai, apibendrinami ir t.t.

Kita duomenų saugyklų svarbi charakteristika, kad **duomenys yra skirtingi laike** (*angl. time variant*). Tai reiškia, kad kiekvienas duomenų vienetas duomenų saugykloje yra tikslus tam tikru laiko momentu. T.y., iš vienos pusės, įrašas turi laiko žymę (*angl. a record is time stamped*), iš kitos pusės, įrašas turi apdorojimo laiką (*angl. a record has a date of transaction*). Bet bet kuriuo atveju, yra tam tikras laikas, žymintis laiko momentą, kada įrašas yra tikslus. Kitu laiku jis ne būtinai yra teisingas, t.y. atitinka tikrovę.

Duomenys nekinta (*angl. non-volatile*) – nauji duomenys įdedami, kaip priedas prie esamų duomenų, o ne pakaitalas. T.y. duomenų saugykla papildoma.

Dimensinis modeliavimas

Pagrindiniai dimensinio modeliavimo terminai yra **faktas** ir **dimensija**.

Faktas – tai verslo veiklos vertinimas, dažniausiai skaitinis ir adityvus (t.y. mes galime sudėti pvz., kiekius į bendrą sumą, t.y. jie yra sudedami), saugomas faktų lentelėje (*angl. a business performance measurement, typically numerical and additive, that is stored in a fact table*). Žr. žemiau pav.

Daily Sales Fact Table
Date Key (FK)
Product Key (FK)
Store Key (FK)
Quantity Sold
Dollar Sales Amuont

Additive – sudedami, semiadditive – sudedami tik tam tikros dimensijos atžvilgiu, nonadditive – nesudedami.

Faktas leidžia įvertinti (pamatuoti) verslo veiklą. Pvz., pardavimus, t.y. kiek prekių per dieną parduota tam tikroje parduotuvėje ir kiek užtai gauta pinigų. Įvertis yra paimamas dimensijų susikirtimo vietoje (pvz., dienos, producto ir parduotuvės). Eilutė lentelėje atitinka tą įvertį. Įvertis yra eilutė faktų lentelėje. Visi įverčiai lentelėje turi vienodą **grudėtumą**, t.y. turime vertinti visus produktus, pvz., mėnesiais, tam tikrame mieste ...

Faktų lentelės išreiškia daug-su-daug ryšius tarp dimensijų dimensiniame modelyje.

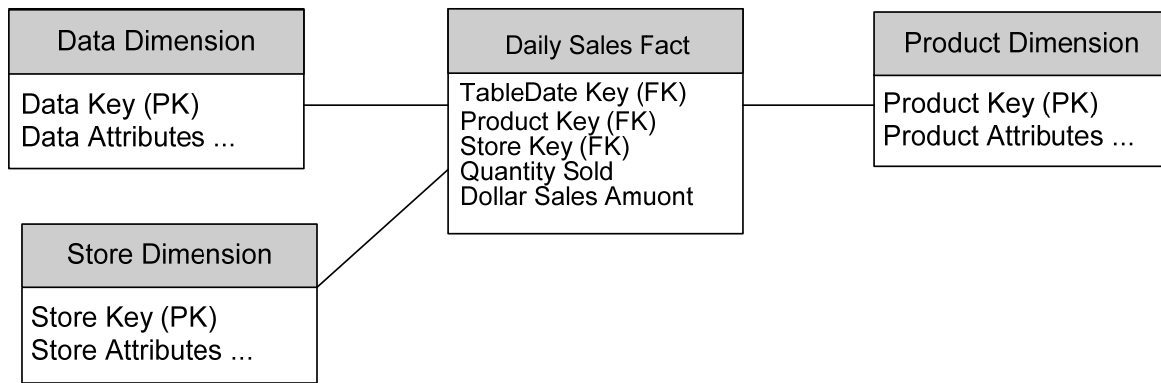
Dimensija – tai nepriklausoma esybė dimensiniame modelyje, kuri yra kaip įeigos taškas arba mechanizmas išrinkti ir sudėti įverčius faktų lentelėje dimensiniame modelyje (*angl. an independent entity in a dimensional model that serves as an entry point or a mechanism for slicing and dicing the additive measures located in the fact table of the dimensional model*).

Dimensijų lentelė yra neatskyriama faktų lentelių dalis. Dimensijų lentelės turinys yra tekstiniai verslo aprašai. Gerai suprojektuotame dimensijų modelyje dimensijų lentelės yra su daug stulpelių arba atributų.

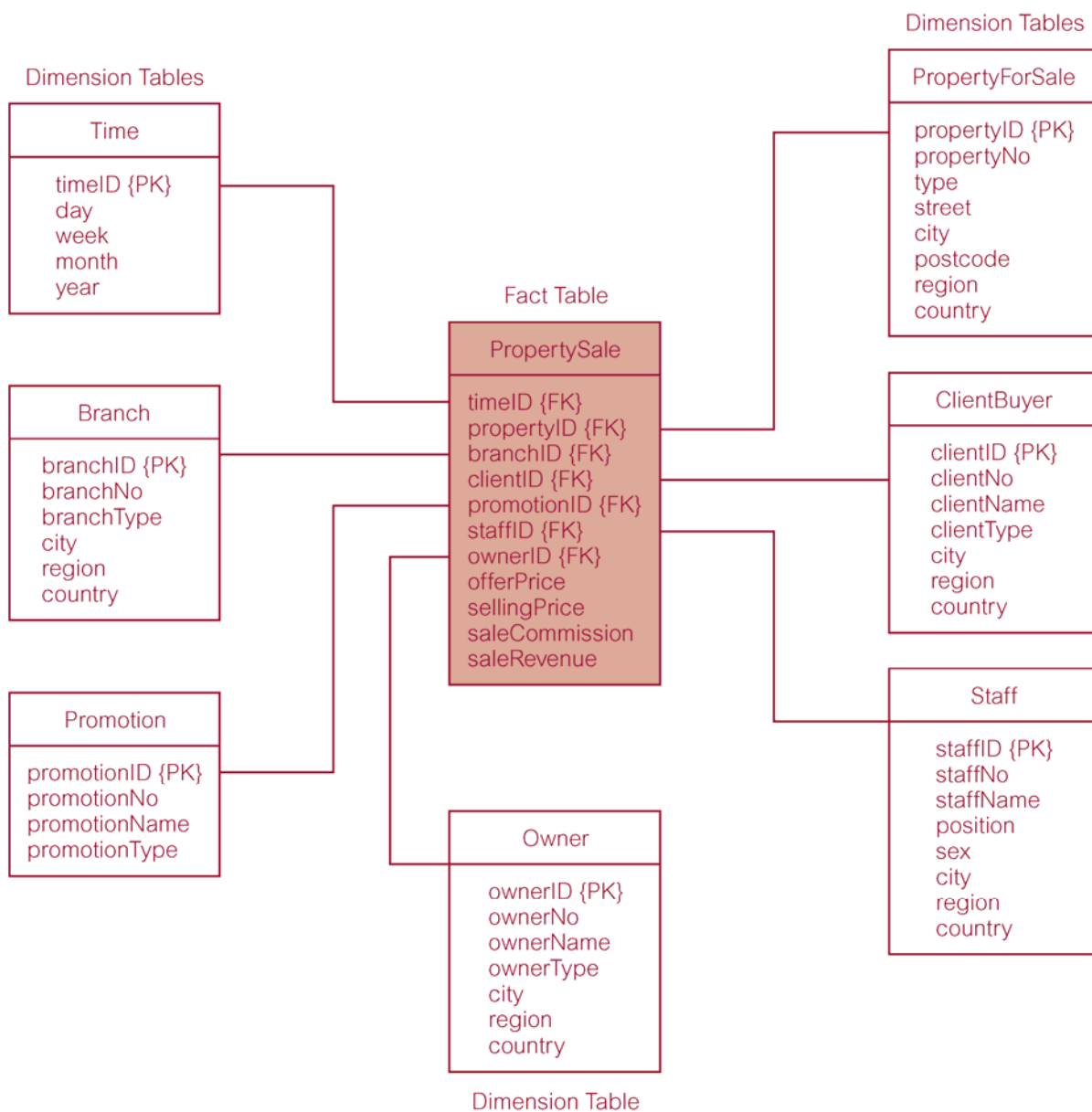
Product Dimension Table
Product Key (PK)
Product Description
SKU Number (Natural Key)
Brand Description (gamintojo)
Category Description
Department Description
Package Type Description
Package Size
Fat Content Description
Diet Type Description
Weight
Weight Units of Measure
Storage Type
Shelf Life Type
Shelf Width
Shelf Height
Shelf Depth
... and many more

Dimensijų lentelės yra kaip įeiga į faktų lentelę. Geriausi atributai yra tekstiniai ir pavieniai (*angl. discrete*). Dydis nors ir skaitinė reikšmė, tačiau naudojama produktui apibūdinti.

Faktų ir dimensijų lentelės yra **dimensinių modelių** sudedamosios dalys. Kaip galima pastebėti, faktų lentelė yra sudaryta iš skaitinių įverčių ir prijungta prie aibės dimensinių lentelių su aprašančiais atributais. Tokia struktūra yra vadinama **žvaigždės schema**. Šita schema pasižymi paprastumu ir lengvai yra suprantama.



Kitas pvz.:

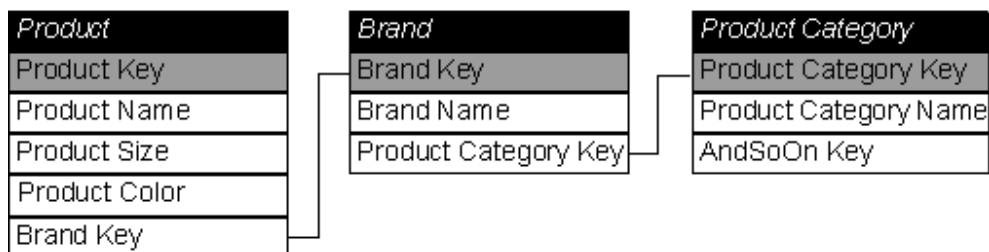


Iš ataskaitų sudarymo pusės, dimensijų atributai yra naudojami, kaip ataskaitos žymės (*angl. label*), o faktų lentelės – kaip skaitmeninės reikšmės ataskaitose.

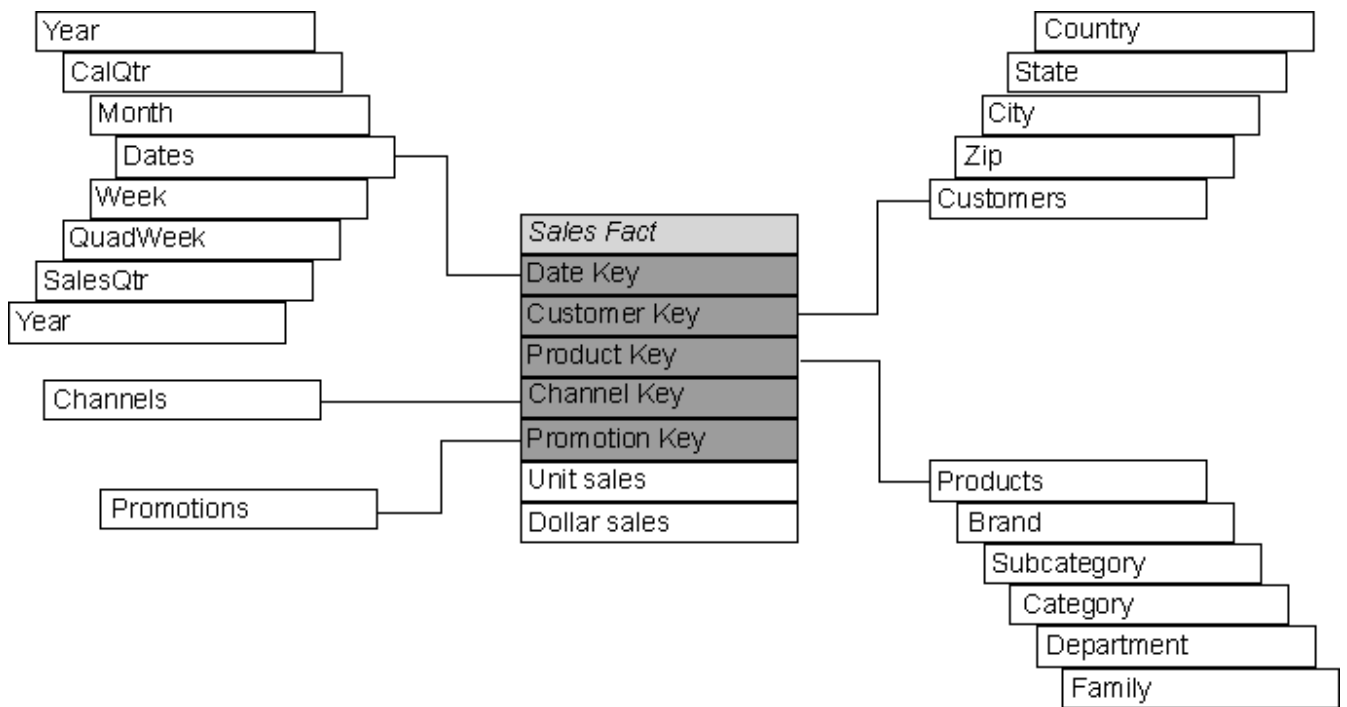
Tai gi galima treipti, kad taip pateikti duomenys yra struktūrizuoti pagal dimensijas. Tačiau, yra tiesioginis ryšys tarp dimensinio ir normalizuoto modelių. T.y. viena ER diagrama dažniausiai yra išskaidoma į daugelis dimensinių modelių. Dideliame normalizuotame įmonės modelyje gali būti pavaizduoti kartu ir pardavimai, ir užsakymai, ir pakrovimų važtaraščiai, ir klientų mokėjimai, ir produktų gražinimai. Tai gi tam tikra prasme, normalizuotuose ER diagramuose yra pavaizduota daugelis verslo procesų (*angl. business process*), kurie niekada neegzistuoja kartu duomenų ir laiko požiūriu. Todėl nieko nuostabaus, kad normalizuoti modeliai atrodo sudėtingi.

Jeigu jau yra ER diagrama, pirmas žingsnis, konvertuojant ją į dimensinį modelį, yra išskaidyti ER diagramą į atskirus verslo procesus ir po to kiekvieną iš jų modeliuoti. Antras žingsnis yra išrinkti daug-su-daug ryšius ER diagramoje, turinčius skaitinius ir adyvičius neraktinius faktus ir pažymėti (*angl. designate*) juos faktų lentele. Paskutinis žingsnis yra denormalizuoti visas likusias lenteles į plokščias lenteles su vienareikšmiais (*angl. single-part*) raktais, kurie yra tiesiogiai sujungiami su faktų lentelėmis. Šitos lentelės tampa dimensinėmis.

Gali būti sudarytos ir **snaigės** (*angl. snowflake*) **tipo schemas**. Schema vadinama snaigės schema, jeigu viena ar daugiau dimensijų lentelių nėra susijusios tiesiogiai su faktų lentele, bet turi būti prijungtos per kitas dimensijų lenteles. Pvz., dimensija, aprašanti produktus, gali būti išskirstyta į tris lenteles (snaigę), kaip pavaizduota žemiau. (*brand – fabrikas*)



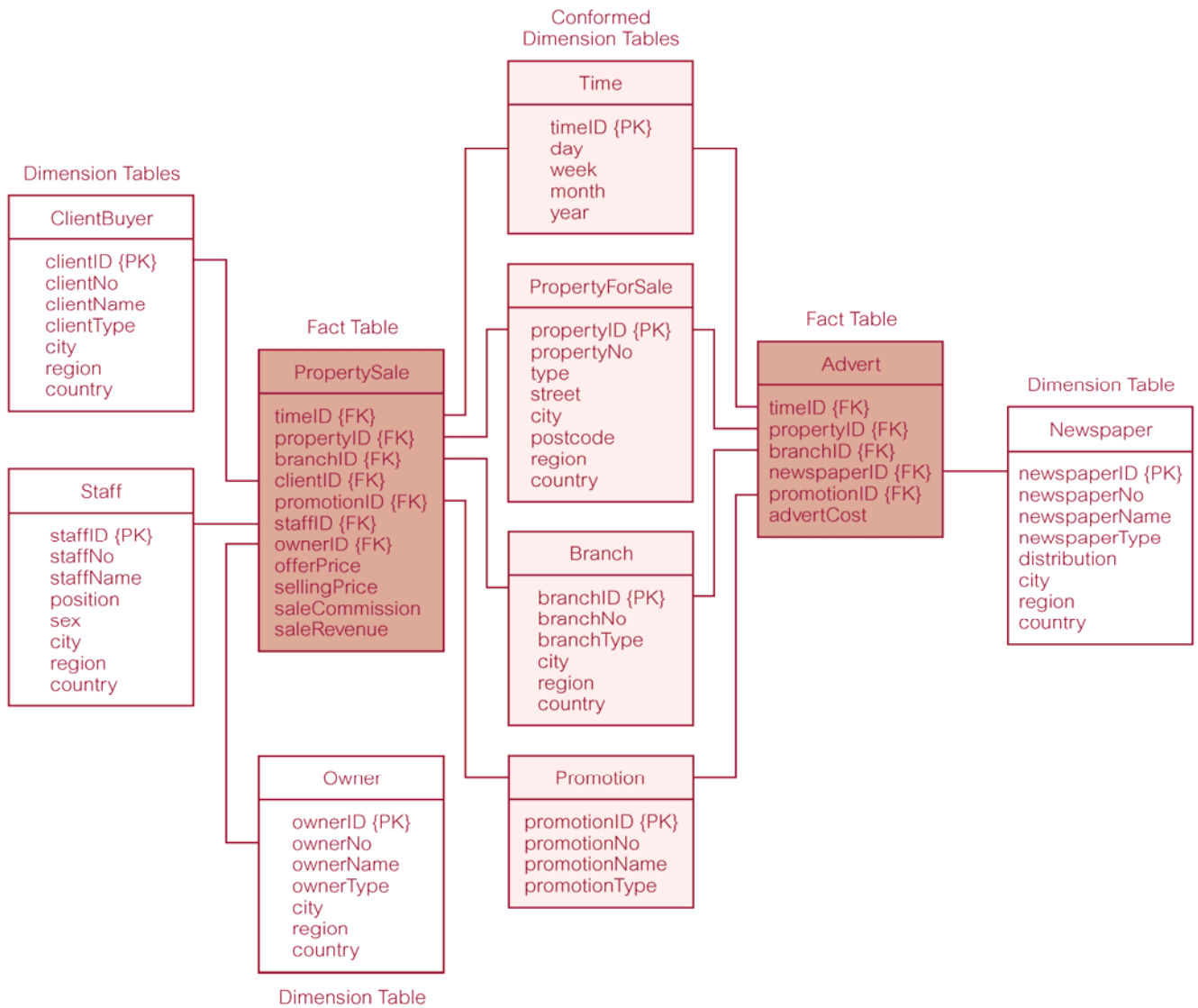
Snaigės schema su daugeliu dimensijų yra pateikta žemiau.

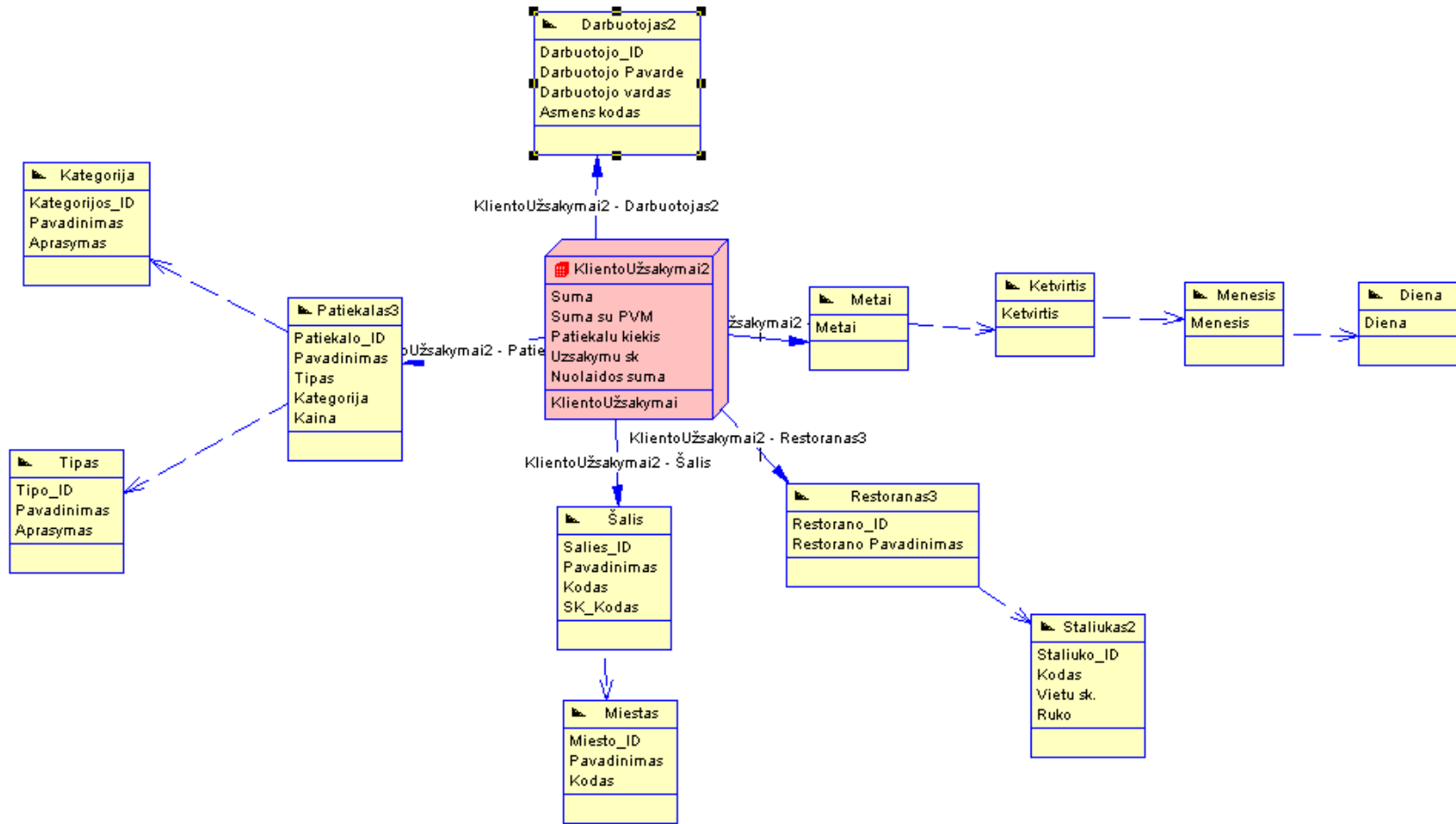


Žvaigždės ar snaigės schema

Abi ir žvaigždės, ir snaigės schemas yra dimensiniai modeliai. Skirtumas yra jų fizinis įgyvendinimas. Snaigės schemas yra lengviau palaikyti, nes jos yra labiau normalizuotos. Žvaigždės schemas yra lengviau prieinamos vartotojams ir dažnai palaiko paprastesnes ir efektyvesnes užklausas (*angl. support simpler and more efficient queries*). Kokią schemą pasirinkti, priklauso nuo pačių dimensijų, t.y., kaip dažnai jos keičiasi ir kokie jų elementai keičiasi. Taip pat reikia pasirinkti tarp to, ar svarbiau lengvas palaikymas ar naudojimas. Lengviausia yra palaikyti sudėtingas dimensijas, pavaizduotas žvaigžde. Išskiriant skirtingus lygmenis į atskiras lenteles, yra užtikrinama ryšio darna (*angl. referential integrity*). Analizės servisai (*angl. Analysis Services*) skaito ir iš snaigės ir iš žvaigždės schemų taip pat gerai. Tačiau, yra svarbu pateikti paprastą ir patrauklų vartotojo interfeisą verslo vartotojams, kurie kuria specialias užklausas dimensinėms duomenų bazėms. Gal būti geriau sukurti žvaigždės versiją snaigės diagramai vartotojams.

Faktų žvaigždynas:





Literatūros saraksts

1. W. H. Inmon. Building the Data Warehouse. Fourth edition. Wiley. 2005.
2. R. Kimball, M. Ross. The Data Warehouse Toolkit. Second edition. Wiley. 2002.
3. Data Warehouse Design Considerations. SQL Server 2000 Resource Kit. URL:
<http://www.microsoft.com/technet/prodtechnol/sql/2000/reskit/part5/c1761.msp>