KAUNO TECHNOLOGIJOS UNIVERSITETAS

INFORMATIKOS FAKULTETAS

PROGRAMŲ INŽINERIJOS KATEDRA

Mantas Kareiva

SKAIČIAVIMŲ, PANAUDOJANT DUOMENŲ KUBUS, ORGANIZAVIMAS IR TYRIMAS

Magistro darbas

Darbo vadovas

prof. E. Bareiša

Kaunas, 2008

KAUNO TECHNOLOGIJOS UNIVERSITETAS

INFORMATIKOS FAKULTETAS

PROGRAMŲ INŽINERIJOS KATEDRA

Mantas Kareiva

SKAIČIAVIMŲ, PANAUDOJANT DUOMENŲ KUBUS, ORGANIZAVIMAS IR TYRIMAS

Magistro darbas

Recenzentas

prof. L. Nemuraitė

Darbo vadovas

prof. E. Bareiša

Atliko

IFM-2/2 gr. stud.

M. Kareiva

2008-05-23

Kaunas, 2008

Turinys

SUMMARY

Data cube precalculation performance related data arrangement and research

Data cube pre computing is time and computer resources consuming task. In spite of this it needs to be done in order to construct an OLAP cube to take advantage of fast querying in data sets enormous in its sizes.

Telecommunication industries collect huge amount of data about events in its networks. Every data portion holds a lot of information (i.e. service type, originator, receiver, time for start, duration, data volume, calling direction, cost, network interface address, etc.). In mobile telecommunication industries it is common to award each customer / subscriber by some prize (money, cell phone, discount to services and so on) in return of 24 month obligation to use one's services. So, every 24 months subscriber gains ability to choose another telecommunication network. In order to maintain stable amount of subscribers' service provider must offer something in return. In order to do that correctly, without financial loses, one must know exact usage statistics of each subscriber.

This paper covers couple tips to arrange data in data warehouses (data marts) in order to achieve greater data cube pre processing speed. One of these methods covers partial data aggregation to highest degree, still sufficient to answer specific queries. Another method shows the ability to synthesize data cube dimensions in order to lower data volumes, that data cube pre calculation could take less time.

PRATARMĖ

Gerinant teikiamų mobiliojo ryšio paslaugų kokybę ir augančius vartotojų poreikius labai svarbu laiku atkreipti dėmesį į kiekvieną mobiliojo ryšio kompanijai įsipareigojusį asmenį bei sudaryti jam palankias sąlygas ir toliau naudotis teikiamomis paslaugomis.

Minėtam tikslui virtualus mobiliojo ryšio operatorius UAB Teledema iškėlė uždavinį sukurti esamų klientų lojalumo administravimo sistemą. Vienas iš sistemos kūrimo tikslų – duomenų, reikalingų nuspręsti kokio dydžio ir kokios formos subsidijas taikyti abonentams, pateikimas. Iš gautų duomenų įmonės marketingo skyrius gali nustatyti tokias lojalumo skatinimo (subsidijavimo) formas, kurios tam tikram laikotarpiui garantuos įmonei pelną. Taip pat iš gautų duomenų tiesiogiai su lojaliais abonentais dirbantys įmonės darbuotojai gali nustatyti kokią subsidijavimo formą taikyti konkrečiam abonentui.

Kadangi apibendrintus duomenis apie klientus reikia pateikti įvairiais detalumo lygmenimis – buvo pasirinkta duomenų agregavimo duomenų kubuose technologija.

Įmonėje dominuoja Microsoft programinė įranga. Ne išimtis ir duomenų bazių serveriai bei programavimo įrankiai. Sistemai kurti buvo naudojama Microsoft Visual Studio 2005 Professional programų kūrimo aplinka su integruotu Microsoft Buisness Intelligence paketu, Microsoft SQL Server 2005 Developer Edition bei Microsoft SQL Server Management Studio (9 v.). Sukurta sistema paskirstyta trijuose serveriuose: Microsoft Windows 2000 Server SBS bei SQL Server 2000, Microsoft Windows 2003 Web Server ir IIS 6.0 bei Microsoft Windows 2003 Server Database Edition ir Microsoft SQL Server 2005.

1. ĮVADAS

Gerinant teikiamų mobiliojo ryšio paslaugų kokybę ir augančius vartotojų poreikius labai svarbu laiku atkreipti dėmesį į kiekvieną mobiliojo ryšio kompanijai įsipareigojusį asmenį bei sudaryti jam palankias sąlygas ir toliau naudotis teikiamomis paslaugomis.

Kadangi mobiliojo ryšio rinkoje konkurencija aštri ne tik rinkodaros ir pardavimų srityje, tačiau ir kainų politikoje – paslaugų kainos yra pasiekusios tokią ribą, kuomet kiekvienas naujas klientas sąlyginai neatneša jokio papildomo pelno mobiliojo ryšio kompanijai, palyginus su išlaidomis, kurios patiriamos tam abonentui privilioti.

Vienintelis būdas pritraukti klientus iš kitų operatorių yra gerinti paslaugų kokybę, o vienintelis būdas neprarasti savų klientų – nuolatos jais rūpintis ir siūlyti optimalias ryšio operatoriui bei klientui paslaugų teikimo, atsiskaitymo, ryšio kokybės bei gero aptarnavimo sąlygas. Kadangi rinkodara ir pardavimai kainuoja daug, tad išlaikyti turimus klientus tampa vis svarbiau.

Klientų išlaikymui skirta rinkodaros dalis yra pigesnė dėl tiesioginio kontakto su klientu, taip pat dėl turimų duomenų apie abonento naudojimosi paslaugomis įpročius. Statistiškai apdorojus ir suintegravus duomenis į iš anksčiau paruoštus ir pagal naudojimosi paslaugomis statistiką suskirstytus pasiūlymus galima paprastai ir greitai pasiūlyti abonentui tai, kas jam yra aktualu, priimtina, pigu, paprasta, pažįstama ir kitaip traukia.

Telekomunikacijų bendrovės operuoja dideliais duomenų srautais. Tiriamo mobiliojo ryšio operatoriaus statistinis abonentas per mėnesį sugeneruoja apie 220 telekomunikacinių įvykių įrašų į duomenų bazes. Vadinasi 1 mln. aktyvių abonentų duomenų bazes papildytų maždaug ketvirčiu milijardo įrašų per mėnesį.

Daugelis mobiliojo ryšio naudojimosi sutarčių sudaromos terminuotam dviejų metų laikotarpiui taikant tam tikras subsidijas už abonentų prisiimtus įsipareigojimus. Besibaigiant laikotarpiui, kuriam buvo įsipareigota naudotis mobiliojo ryšio operatoriaus paslaugomis, apie kiekvieną abonentą operatorius turi prikaupęs vidutiniškai 5250 įrašų duomenų bazėje. Jei vienu metu reikia suskaičiuoti 100 000 abonentų naudojimosi statistiką už paskutinius du metus – vien naudingų įrašų skaičiuojanti sistema apdoroja pusę milijardo. O tuos įrašus reikia atrinkti iš bendro įrašų kiekio, kuris būna gerokai didesnis.

Apibendrinti statistikos apie visų abonentų naudojimąsi paslaugomis rezultatai taikomi sprendžiant apie galimas suteikti mobiliojo ryšio operatoriui pelningas subsidijas norint išlaikyti turimą abonentą. Konkretūs abonento naudojimosi paslaugomis statistikos rezultatai taikomi sprendžiant kokią subsidiją (iš galimų) taikyti konkrečiam abonentui. Taigi mobiliojo ryšio operatoriui reikia turėti ir darbuotojams operatyviai pateikti minėtą informaciją.

Kadangi duomenų agregavimas naudojant SQL kalbos operatorių GROUP BY yra vienmatis – jo naudojimas duotu atveju yra nepakankamas. Skaičiavimų rezultatai turi būti pateikiami N-matėje erdvėje (Grey, Bosworth ir Layman). Šiuo konkrečiu atveju erdvės matiškumą nusako diskretiniai laiko intervalai (diena, mėnuo, metai, įsipareigojimo laikotarpis) bei abonentų skirstymas į mobiliojo ryšio rinkai būdingas sritis (mokėjimo planas, kontraktas, abonentas).

Daugiamatėms duomenų agregavimo operacijoms atlikti ir rezultatams saugoti bei pateikti skirti duomenų kubai. Duomenų kubas – tai daugiamatė duomenų struktūra, literatūroje kartais vadinama daugiamačiu masyvu (Grey, Bosworth ir Layman), (Vilimas).

Dideliems duomenų kubams apskaičiuoti išnaudojama daug kompiuterinio laiko. Tačiau, priklausomai nuo reikiamų gauti rezultatų, kai kurių skaičiavimų trukmę galima ženkliai sumažinti. Dauguma autorių (Barbará ir Sullivan), (Dehne, Eavis ir Hambrusch), (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers), (Lee, Ling ir Li), (Shanmugasundaram, Fayyad ir Bradley), (Vitter, Wang ir Iyer), (Wang, Lu ir Feng) siūlo įvairių kubo skaičiavimo greitinimo algoritmų.

1.1. Dokumento paskirtis

Šio darbo tiriamasis objektas – duomenų kubo pirminio skaičiavimo spartinimas. Toliau darbe aptariama ar, priklausomai nuo reikiamo duomenų pateikimo būdo bei duomenų prigimties, galima sumažinti apdorojamo kubo dydį. Keliama hipotezė, kad tai, dėl sumažėjusio įrašymo / skaitymo (*I/O*) operacijų kiekio, turėtų paspartinti pirminį kubo skaičiavimą.

Darbą sudaro pratarmė, santrauka anglų kalba, įvadas, 3 skyriai, išvados ir pasiūlymai, naudotos literatūros sąrašas, priedai (kompaktinė plokštelė su eksperimentų vykdymo trukmių rezultatais; projekto analogų analizė; sistemos įdiegimo įmonėje aktas).

Darbo apimtis - 63 puslapiai, jame yra 4 lentelės ir 22 paveikslai. Literatūros sąrašą sudaro 24 šaltiniai. Darbo pabaigoje pateikti 3 priedai.

1.2. Santrauka

Duomenų kubo konstravimas yra laikui ir kompiuteriniams resursams imlus procesas. Nepaisant to, šis darbas turi būti atliktas norint pasinaudoti greitų užklausų iš ypatingai didelių OLAP kubų teikiamais privalumais .

Telekomunikacijų bendrovės surenka didelius duomenų kiekius apie įvykius telekomunikaciniuose tinkluose. Kiekviena duomenų porcija aprašo daug informacijos (pavyzdžiui: paslaugos tipą, iniciatorių, gavėją, pradžios laiką, trukmę, perduotų duomenų kiekį, skambučio kryptį, kainą, tinklo sąsajos adresą ir t.t.). Mobiliojo ryšio rinkoje yra įprasta apdovanoti kiekvieną abonentą tam tikru prizu (pinigais, nuolaidomis ar nauju mobiliuoju telefonu) mainais į 24 mėnesių sutartį naudotis konkretaus operatoriaus paslaugomis. Taigi kas 24 mėnesius abonentas turi galimybę pakeisti paslaugos teikėją. Tam, kad ryšio operatorius išlaikytų savo klientus, už sutarties pratęsimą taip pat turi pasiūlyti dovaną. Kad būtų galima tai atlikti nepatiriant finansinių nuostolių – mobiliojo ryšio operatorius privalo žinoti kiekvieno abonento naudojimosi paslaugomis statistiką.

Šiame dokumente aprašoma pora būtų kaip pakeisti duomenų pirminį vaizdą (struktūrą ir sudėtį) siekiant pagreitinti duomenų kubų konstravimo procesą. Vienas šių metodų – duomenų agregavimas iki didžiausio, vis dar tinkamo analizei, lygio. Kitas metodas – tai lėtai kintančių kubo dimensijų sintezavimas taip sumažinant kubo dydį ir pagreitinant jo kūrimą.

2. DUOMENŲ KUBŲ PIRMINIO SKAIČIAVIMO PASPARTINIMO GALIMYBIŲ ANALIZĖ

Šiame skyriuje aprašomas tyrimo sritis ir objektas, problematika ir galimi teoriniai jos sprendimo būdai.

2.1. Sistemos kūrimo tikslas

Sistemos kūrimo tikslas yra paprastinti mobiliojo ryšio paslaugų teikėjo abonentų, kurių sutartis artėją į pabaigą, sutarties pratesimo administravimą bei sumažinti rinkodaros, skirtos minėtai abonentų grupei, kaštus.

2.2. Tyrimo tikslas

Tyrimo sritis: duomenų kubų pirminiai skaičiavimai (*precomputing*).

Tyrimo objektas: Microsoft SQL Server 2005 Analysis Services duomenų kubai.

Tyrimo kontekstas: mobiliojo ryšio operatoriaus kaupiama klientų naudojimosi paslaugomis statistika.

Problema: suformuoti duomenų kubą (atlikti pirminius duomenų kubo skaičiavimus, kitaip – kubo konstravimą) iš didelio duomenų kiekio užtrunka reikšmingą laiko tarpą. Net ir naudojant modernią techninę įrangą kai kurių dalykinių sričių specifika lemia keleto parų skaičiavimų trukmę.

Šiame darbe apžvelgiame duomenų kubo pirminio skaičiavimo (konstravimo) trukmę konkrečiai dalykinei sričiai sumažinusį duomenų paruošimo būdą.

2.3. Esminės projekto įgyvendinimo problemos

Kuriant mobiliojo ryšio operatoriaus abonentų lojalumo valdymo sistemą susiduriama su daugeliu jos įgyvendinimo kliūčių. Vienas iš didžiausių sunkumų yra

integracija į jau egzistuojančios sistemos aplinką ir nuolatinis stabilus darbas naudojant turimus techninius resursus.

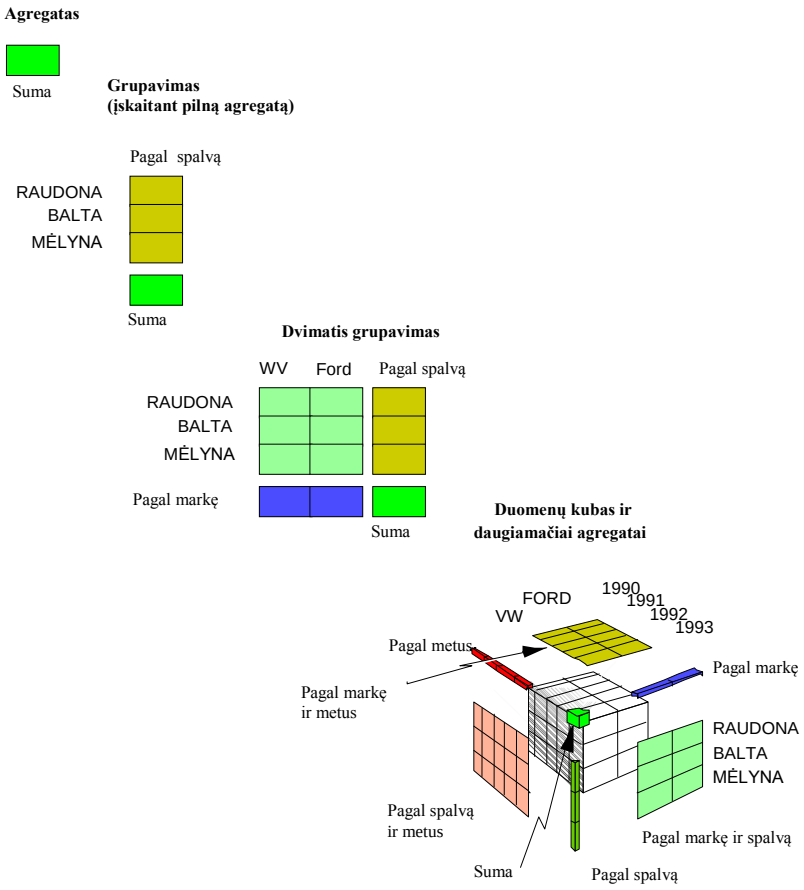
Kadangi duomenų kubų pirminiai skaičiavimai yra imlus techniniams resursams procesas, tad panaudojant šio darbo rezultatus siekiama minimizuoti techninės įrangos apkrovimą atliekant duomenų analizes.

Nors konkrečiu momentu pateikiama abonento naudojimosi paslaugomis statistika galutinėje išraiškoje yra nedidelės apimties, tačiau jai apskaičiuoti panaudojama nemažai kompiuterio darbo laiko. Išsamiai problema aprašyta 2.5 Duomenų apdorojimo trukmės problema skyriuje po to, kai skaitytojas supažindinamas su duomenų kubų specifika.

2.4. Duomenų kubo apibrėžimas

Duomenų analizės programinė įranga paprastai agreguoja daugelio matmenų duomenis, ieškodama anomalijos ar neįprastos duomenų struktūros. SQL agregavimo funkcijų ir GROUP-BY operatoriaus rezultatas – vienmatis arba be matmenų (kuomet grupavimo požymis nėra nurodomas). Programos, agregavimui naudojančios duomenų kubus, rezultatus apibendrina naudodamos sudėtingesnius ryšius tarp duomenų, nei tik grupavimas. Naudojami tarpusavio ryšiai tarp duomenų lentelės stulpelių, dalinis sumavimas, paieška į gylį (Grey, Bosworth ir Layman). Tad apibendrinus galima pasakyti, kad duomenų kubas yra daugiamatė duomenų struktūra, kurios

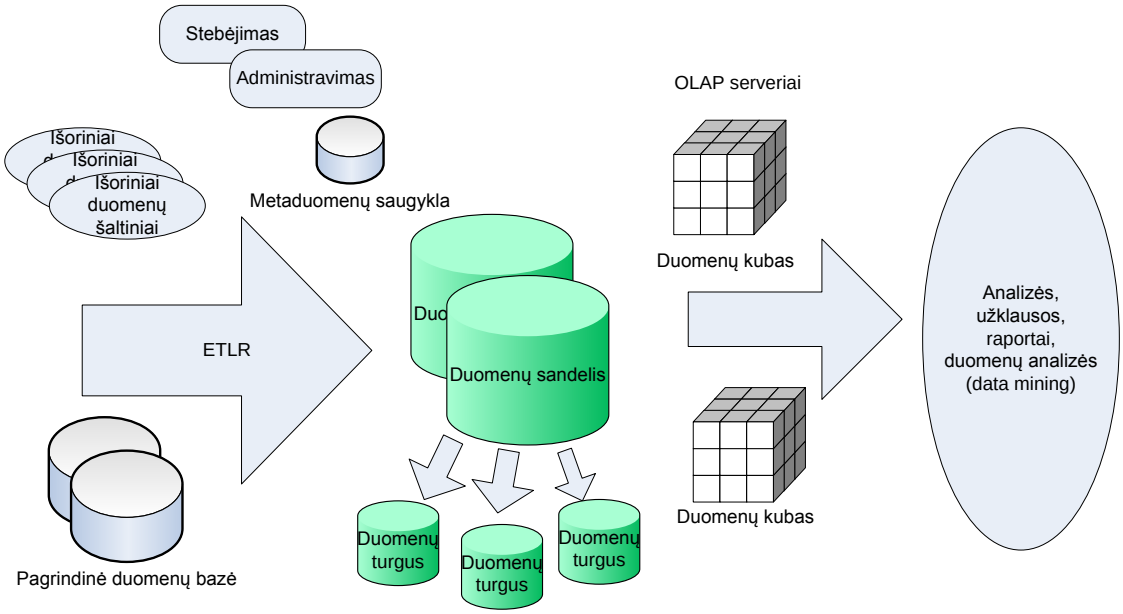
- matmenys – tai grupavimo operatoriai,
- viduje yra faktų (įvykių, duomenų porcijų) sąrašas;
- briaunose yra faktų sumų (ar kitaip apibendrintų faktų) sąrašas pagal vieną dimensiją;
- kampuose yra faktų sumų (ar kitaip apibendrintų faktų) sąrašas pagal daugiau, nei vieną dimensiją.



Pav. 1 Grafinis trijų matmenų duomenų kubo vaizdas

Šiuolaikinė programinė įranga, skirta duomenų analizei įmonės mastu ir sprendimų priėmimo palengvinimui, dažnai apima tokias paslaugas ir sistemas, kaip įvairūs duomenų šaltiniai (kaip OLTP serveriai), ETL servisai, administravimo ir stebėjimo įrankiai, metaduomenų saugykla, duomenų sandėlis (*data warehouse*), duomenų turgus (*data mart*), OLAP serveriai ir įvairūs įrankiai, skirti analizėms, raportavimui, duomenų užklausoms, duomenų analizėms (*data mining*).

Į įmonės veiklą integruoto duomenų saugojimo ir analizės įrankių sistemos principinė schema pavaizduota Pav. 2 (Chaudhuri ir Dayal), (Pedersen ir Jensen).



Pav. 2 Įmonės sprendimų priėmimo pagalbinės programinės įrangos principinė schema

Matome, kad norimam rezultatui pasiekti – sprendimų priėmimo palengvinimo uždaviniui paprastinti, naudojama platus įmonės informacinių technologijų rinkinys.

2.5. Duomenų apdorojimo trukmės problema

Į apskaitos sistemos duomenų bazę per vieną kalendorinę parą vidutiniškai įrašoma apie 1 200 000 unikalių įrašų. Kiekviename įraše pateikiama tokia informacija:

- Įrašo šaltinis (GSM / UMTS tinklo ar kt. įrenginys, generuojantis telekomunikacinius įvykius);
- Unikalus įrašo identifikatorius šaltinio ribose;
- Įrašo tipas (išeinantis / įeinantis skambutis, SMS, duomenų perdavimas ir kt.);
- Telekomunikacinio įvykio pradžia;
- Telekomunikacinio įvykio pabaiga;
- Telekomunikacinio įvykio trukmė;
- Telekomunikacinio įvykio metu perduotų duomenų kiekis;
- Telekomunikacinio įvykio iniciatoriaus identifikatorius (MSISDN);
- Telekomunikacinio įvykio iniciatoriaus identifikatorius (IMSI);
- Telekomunikacinio įvykio adresatas normalizuotoje formoje;
- Telekomunikacinio įvykio adresatas originalioje formoje;

- Prieigos taško vardas;
- Tinklų sujungimo paslaugos teikėjo vardas;
- SMSC;
- Peradresuojantis numeris;
- Tarptinklinio ryšio operatoriaus kodas;
- Tarptinklinio ryšio įvykio kaina;
- Tarptinklinio ryšio metu suteikto laikinojo nacionalinio MSISDN informacija.

Vieno įrašo vidutinis dydis yra 155 baitai. Taigi per vieną parą duomenų bazė, neskaitant indeksų užimamų duomenų, papildoma 180 MB. Per vieną mėnesį apskaitos duomenų bazė papildoma vidutiniškai 5,5 GB duomenų.

Perkeliant duomenis į duomenų sandėlį (*Data Warehouse*) kiekviena duomenų eilutė papildoma informacija apie telekomunikacinio įvykio kainą, paros laiko skirstymą, papildomai apskaitomų paslaugų informacija tačiau ir jos pašalinama vėlesniuose analizės etapuose nenaudinga informacija apie šaltinio identifikatorių, APN ir pan.

Per mėnesį duomenų sandėlis papildomas apytiksliai 4 GB duomenų apie telekomunikacinius įvykius, paruoštų sąskaitų klientams generavimui ir analizėms. Taigi analizuojant duomenis už 24 paskutinius mėnesius reikia apdoroti apie 95 GB skaitmeninės informacijos.

2.5.1. Duomenų kubo dydžio apskaičiavimas

Pilno duomenų kubo konstravimui iš duomenų bazės nuskaitomas didelis faktų kiekis. Faktai jungiami su dimensijomis pagal kiekvieną dimensijos komponentą. Taip gaunamas kubas, kurio dydis yra:

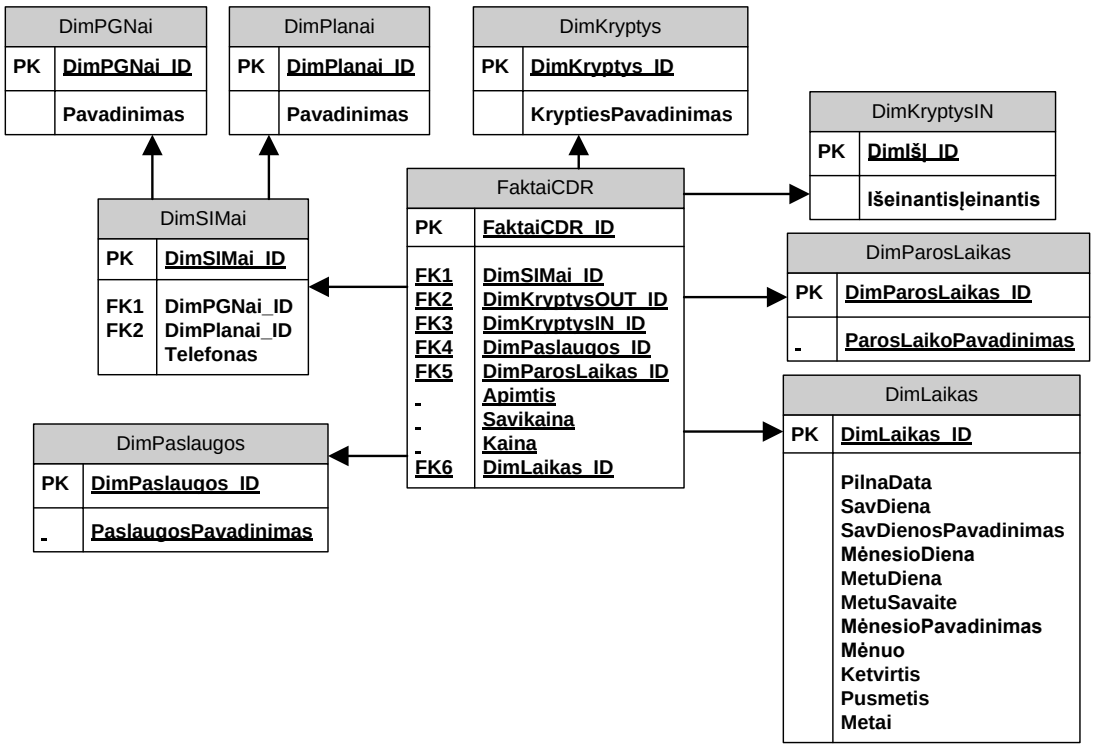
$$n_{kubo} = n_{faktų} \times \sum_{i} n_{dim} \qquad (1)$$

Čia i –dimensijos numeris, n_{kubo} – kubo dydis ląstelėmis, $n_{faktų}$ – faktų kiekis.

2.5.2. Duomenų kubo konstravimas iš duomenų sandėlio

Atliekamos duomenų analizės tikslas yra sužinoti:

1. Kokiu mokėjimo planu besinaudojantis,
2. Kuris konkrečiai abonentas,
3. Į kokius viešojo telefoninio ryšio tinklus (iš kokių viešojo telefoninio ryšio tinklų),
4. Kiek
    a. minučių skambino,
    b. siuntė SMS pranešimų,
    c. siuntė MMS pranešimų,
    d. duomenų persiuntė naudodamasis duomenų perdavimo paslauga,
    e. atitinkamų telekomunikacinių įvykių buvo inicijuota ar terminuota naudojantis tarptinkliniu ryšiu
5. Kiek
    a. sulaukė skambučių minučių
    b. sulaukė SMS
    c. sulaukė MMS
6. Kaip minėti telekomunikaciniai įvykiai pasiskirstę paros laike: kiek
    a. iš jų inicijuota / terminuota piko metu,
    b. iš jų inicijuota / terminuota ne piko metu,
    c. iš jų inicijuota / terminuota neskirstant paros laiko
7. Kokia buvo minėtų telekomunikacinių įvykių savikaina duotuoju momentu,
8. Kokia telekomunikacinių įvykių kaina buvo pritaikyta duotuoju momentu,
9. Vidutiniškai per vieną kalendorinį mėnesį.



Pav. 3 Išsamaus duomenų kubo konstravimui skirta duomenų sandėlio schema

Konstruojant duomenų kubą 1-6, 9 aukščiau paminėti objektai buvo traktuojami kaip kubo matmenys (*dimensions*). 7 ir 8 objektai – kaip matavimai (*measures*). Kadangi lietuvių kalboje terminai *matmenys* ir *matavimai* semantiškai ir akustiškai yra panašūs, tad aiškumo dėlei matmenys toliau šiame darbe vadinami *dimensijomis*.

Dimensijos DimPGNai ir DimPlanai yra vienareikšmiškai nusakomos sutartyje su abonentu ir nekinta sutarties galiojimo laikotarpiu (2 metus). Taigi šios dimensijos tiesioginės įtakos duomenų kubo pirminio skaičiavimo trukmei neturi.

Duomenų kubo suskaičiavimas pagal aukščiau pateiktą dimensijų ir matavimų schemą, pavaizduotą Pav. 3, trunka apie 46 valandas.

2.5.3. Duomenų kubo konstravimas iš duomenų turgaus (data mart)

Hipotezė 1: atlikus tarpinį duomenų agregavimą ir taip sumažinus duomenų šaltinio apimtį – sumažės ir duomenų kubo pirminio skaičiavimo trukmė.

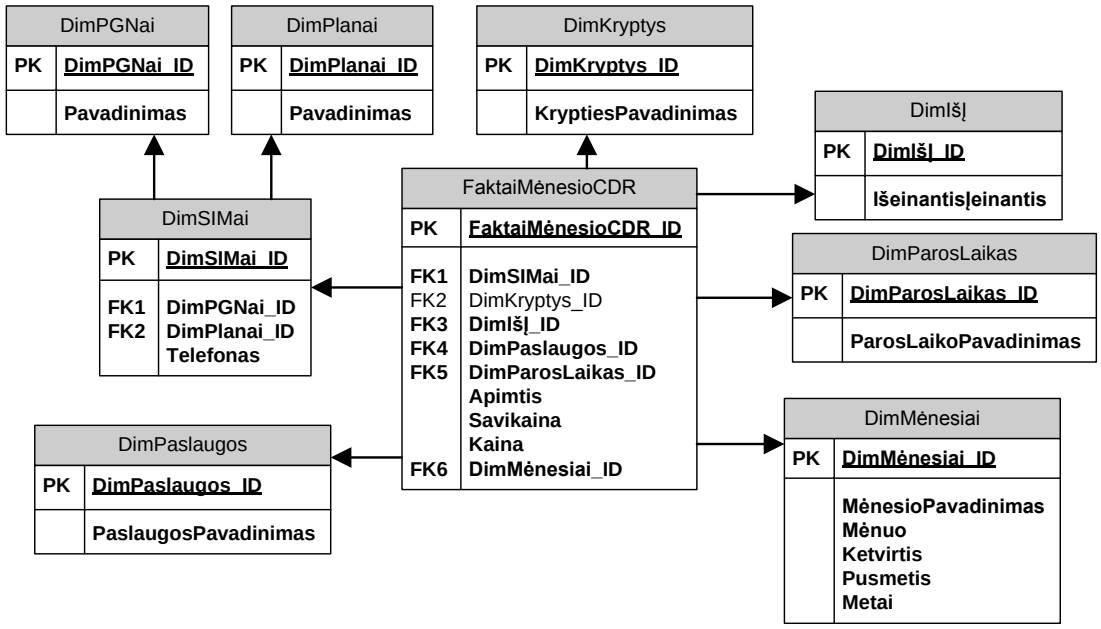
Duomenų turgus (*data mart*) – tai duomenų sandėlio dalis, skirta konkrečiam vartotojų ratui (dažniausiai – konkrečiam įmonės skyriui) pateikti konkrečias ataskaitas.

Dažniausiai duomenų turgus yra paprastą duomenų struktūrą įkūnijanti snaigės principu sujungta duomenų schema.

Kadangi skaičiavimų trukmė, duomenų šaltiniu naudojant duomenų sandėlį, buvo gana didelė, tad nusprendžiau sukonstruoti duomenų turgų.

Iš duomenų sandėlio, kurio struktūra aptarta 2.5.2 Duomenų kubo konstravimas iš duomenų sandėlio skyriuje, sukonstruotas duomenų turgaus esminis skirtumas – tai faktų agregavimas, neprarandant naudingos informacijos.

Pagal dalykinę sritį (žr. skyrių 2.5.2 Duomenų kubo konstravimas iš duomenų sandėlio) tampa aišku, kad pagrindinis pirminio duomenų agregavimo požymis yra 9-tuoju numeriu įvardinta dimensija. Kadangi aptariamoji dimensija tiesiogiai apibrėžia duomenų analizės rezultatui reikalingą formą – vidutinę (2.5.2 Duomenų kubo konstravimas iš duomenų sandėlio) minėtų matavimų reikšmę per mėnesį, tad logiškai galima manyti, jog neprarandant reikiamos informacijos duomenis analizei galima saugoti agreguotoje formoje.



Pav. 4 Duomenų kubo konstravimui skirta duomenų turgaus schema

Pagal Pav. 4 pateiktos schemos semantinę prasmę matome, kad dimensijų lentelė *DimLaikas* pakeista į dimensijų lentelę *DimMėnesiai* bei faktų lentelėje nuoroda į konkretų laiko momentą pakeista nuoroda į vieno kalendorinio mėnesio diskretiškumu laiką aprašančius įrašus. Iš šio duomenų šaltinio sukonstruota duomenų kubą vadinsime **minimaliu duomenų kubu**.

Minimalus duomenų kubas – tai minimalaus detalumo duomenų kubas, reikalingas sudaryti vieno abonento vieno mėnesio naudojimosi paslaugomis statistikai.

2.5.4. Duomenų turgaus faktų lentelės užimamos atminties minimizavimas

Hipotezė 2: duomenų šaltinio struktūrą pakeitus kompaktiškesne dėl sumažėjusio skaitymo / rašymo (*I/O*) operacijų kiekio duomenų kubo pirminio skaičiavimo trukmė turėtų sumažėti.

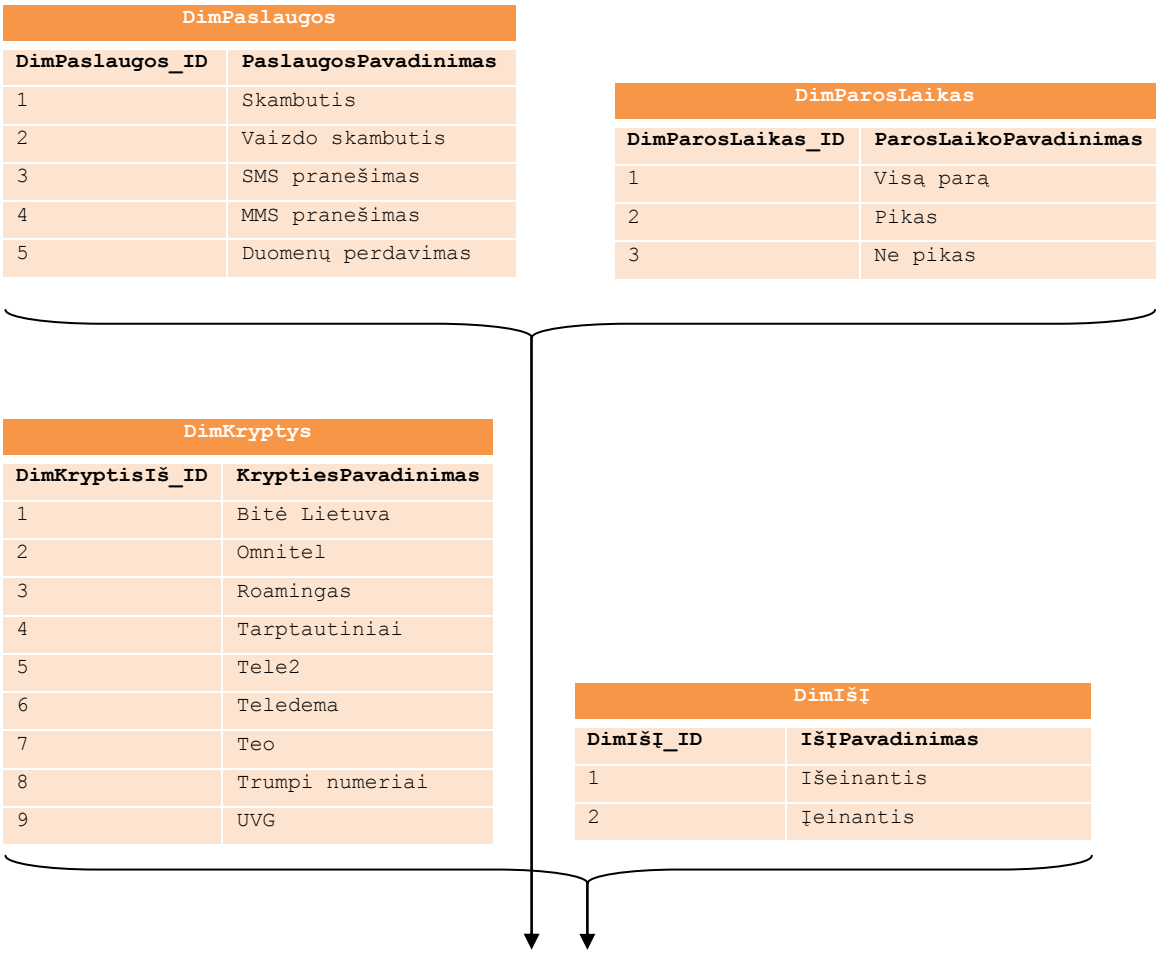
Pagal (Kimball) pateikiamą teoriją ir jos sąsają su dalykine sritimi galime išskirti lėtai kintančias dimensijas ir iš jų dirbtinai sukurti vieną dimensiją.

Pavadinkime ją **sintetine dimensija**.

Dalykinėje srityje lėtai kintančios dimensijos ir jų egzempliorių pavyzdžiai yra:

- Paslaugos
  - Skambučiai;
  - SMS pranešimai;
  - MMS pranešimai;
  - Duomenų perdavimo paslauga
- Kryptys
  - Omnitel;
  - Bitė Lietuva;
  - ...
  - trumpieji numeriai;
  - užsienis
- Įeinantis srautas / išeinantis srautas
  - Įeinantis;
  - Išeinantis
- Paros laikas
  - visą parą;
  - pikas;
  - ne pikas.

Grafiškai pavaizduota dimensijų sintezė atrodo taip:



DimPaslaugos

DimPaslaugos_ID	PaslaugosPavadinimas
1	Skambutis
2	Vaizdo skambutis
3	SMS pranešimas
4	MMS pranešimas
5	Duomenų perdavimas

DimParosLaikas

DimParosLaikas_ID	ParosLaikoPavadinimas
1	Visą parą
2	Pikas
3	Ne pikas

DimKryptys

DimKryptisIš_ID	KryptiesPavadinimas
1	Bitė Lietuva
2	Omnitel
3	Roamingas
4	Tarptautiniai
5	Tele2
6	Teledema
7	Teo
8	Trumpi numeriai
9	UVG

DimIšĮ

DimIšĮ_ID	IšĮPavadinimas
1	Išeinantis
2	Įeinantis

DimSintetinė

DimSintetinė_ID	Reikšmė
1	Išeinantis skambutis piko metu į Bitę
2	Įeinantis skambutis piko metu iš Bitės
3	Išeinantis skambutis ne piko metu į Bitę
4	Įeinantis skambutis ne piko metu iš Bitės
5	Išeinantis skambutis visą parą į Bitę
6	Įeinantis skambutis visą parą iš Bitės
7	Išeinantis SMS piko metu į Bitę
8	Įeinantis SMS piko metu į Bitę
9	Išeinantis SMS piko metu į Bitę
...	...
269	Išeinantis duomenų perdavimas visą parą į UVG
270	Įeinantis duomenų perdavimas visą parą iš UVG

Matome, kad sintetinės dimensijos galimų reikšmių kiekis yra išreiškiamas formule

$$\forall D \in DS, sintD\,(n) = \sum D_i(n) \qquad (2)$$

Čia *D* – dimensija, *DS* – salyklinė sritis, *n* – dimensijos unikalių reikšmių kiekis, *sintD* – sintetinė dimensija.

Priklausomai nuo dalykinės srities kai kurios sintetinės dimensijos narių reikšmės yra nelogiškos. Kadangi duomenų perdavimas yra betarpiškai vykdomas tarp konkrečios SIM kortelės ir APN, tad sintetinės dimensijos reikšmės „*Išeinantis duomenų perdavimas visą parą į UVG*“ yra nenaudotinas. Dimensijos „*DimPaslaugos*“ reikšmė „*Duomenų perdavimas*“ gali vienareikšmiškai būti sintetinama tik su „*DimParosLaikas*“ ir „*DimIšĮ*“ dimensijomis. Su dimensija „*DimKryptys*“ „*Duomenų perdavimas*“ reikšmę logiška sintetinti tik semantiškai atrinkus „*Roamingas*“ reikšmę ir apibendrinus „*Duomenų perdavimas*“ pagal kitas dimensijas. Tą parodo ir neprivalomas *Kryptys_ID* parametras duomenų turgaus duomenų struktūroje (žr. Pav. 4 Duomenų kubo konstravimui skirta duomenų turgaus schema).

Iš paminėtų faktų seka, kad automatinis sintetinių dimensijų generavimas ne visuomet tikslingas. Tačiau, įvertinant sukurtosios sintetinės dimensijos apimtį ir žmogaus darbo laiką, reikiamą atlikti semantinei duomenų turgaus analizei, matome, kad automatinis generavimas taip pat gali būti reikšmingas laiko taupymo prasme kuriant duomenų struktūrą.

Tačiau net automatiškai sugeneravus pilną sintetinę dimensiją joje truktų semantiškai logiškų reikšmių (pvz.: „*Išeinantis duomenų perdavimas visą parą*“ ir „*Išeinantis duomenų perdavimas visą parą roamingas*“ bei atitinkamų reikšmių pagal kitas dimensijas). Taigi automatiniam sintetinės dimensijos generavimui reikalingas algoritmas, tikrinantis ar ryšiai tarp konkrečios fakto reikšmės ir konkrečios dimensijos yra privalomi.

Taigi formulę (2) norint panaudoti automatiniam generavimui reikėtų pertvarkyti taip:

$$\forall D \in DS, sintD\,(n) = \sum D_i(n) + \sum Dp_i \qquad (3)$$

Čia Dp_i – tai tokios dimensijos, kurios su faktais susijusios neprivalomais ryšiais, o išraiška $\sum Dp_i$ reiškia tokių dimensijų kiekį.

Pagal formulę (3) būtų galima automatiškai sugeneruoti pilną **perteklinę sintetinę dimensiją**.

Pertekline sintetine dimensija vadinkime tokią sintetinę dimensiją, kurioje yra visų lėtai kintančių dimensijų kortežų CROSS JOIN operacijos rezultatas bei, papildomai, CROSS JOIN rezultatas tų dimensijų, kurios susijusios privalomais ryšiais su duomenų šaltiniu (faktais).

Eksperimento metu (parodyta vėlesniuose šio dokumento skyriuose) buvo nustatyta, kad sintetinės dimensijos perteklius žymios įtakos kubo pirminio skaičiavimo greičiui neturi jei dimensijos užimama vieta atmintyje yra sąlyginai nedidelė turint omenyje techninės įrangos galimybes.

Ši duomenų kubų užimamos atminties mažinimo metodika turėtų daryti teigiamą įtaką pirminio kubo skaičiavimui. Įtaka turėtų pasireikšti dėl sumažėjusio operatyvinės atminties poreikio ir mažesnio rašymo / skaitymo operacijų į / iš išorinės kompiuterio atminties kiekio.

Kaip pasireiškia tokio dimensijų sintetinimo nauda parodoma 4.2 skyriuje.

2.6. Kiti galimi naudoti kubo skaičiavimo spartinimo metodai

Kai kurie autoriai ((Barbará ir Sullivan), (Dehne, Eavis ir Hambrusch), (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers), (Lee, Ling ir Li), (Shanmugasundaram, Fayyad ir Bradley), (Vitter, Wang ir Iyer), (Wang, Lu ir Feng)), siūlo įvairias kubų skaičiavimo trukmės mažinimo metodikas. Kubų skaičiavimo greitį galima padidinti

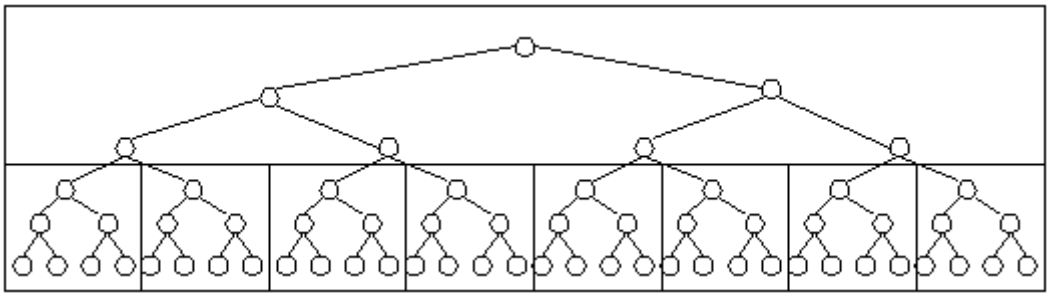
- atitinkamai indeksuojant duomenų šaltinius (Microsoft), (Baniulis, Ds2005/Teorija), (Baniulis, Bajerio medis);
- paralelizuojant skaičiavimus (Dehne, Eavis ir Hambrusch), (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers), (Goil ir Choudhary, A parallel scalable infrastructure for OLAP and data mining);
- pritaikant specifinius atminties valdymo mechanizmus (Agarwal, Agrawal ir Deshpande),(Dasu, Johnson ir Baker);
- mažinant rezultatų tikslumą (Vitter, Wang ir Iyer), (Shanmugasundaram, Fayyad ir Bradley), (Barbará ir Sullivan);
- mažinant kubų dydį (Wang, Lu ir Feng), (Lee, Ling ir Li).

2.6.1. Duomenų šaltinio indeksavimas

(Baniulis, Bajerio medis) Viena paprasčiausių ir praktikoje dažnai naudojamų medžio tipo struktūrų yra binarinis arba dvejetainis medis. Bet dažnai reikia medžio, kurio kiekviena viršūnė turėtų daugiau negu dvi šakas. Tokie medžiai vadina daugiašakiais medžiais.

Viena iš pagrindinių daugiašakių medžių panaudojimo sričių – tai daugiaelementinių paieškos medžių formavimas ir jo funkcionavimo palaikymas. <...> Pagrindinė problema, su kuria susiduria duomenų bankų kūrėjai yra algoritmo efektyvumas paieškos grečio bei naudojamos atminties atžvilgiu.

Daugiašakiai medžiai idealūs tokių uždavinių sprendimui, nes jei mes kreipiamės į vieną elementą, kuris saugojamas atmintyje, tai be jokių papildomų sąnaudų mes galime kreiptis ir į tam tikrą grupę elementų. Tai reiškia, kad medis suskaidytas į dalinius medžius**,** ir daliniai medžiai ir sudaro elementų grupes, kurios prieinamos vienu metu. <...> tokie daliniai medžiai vadinami puslapiais.

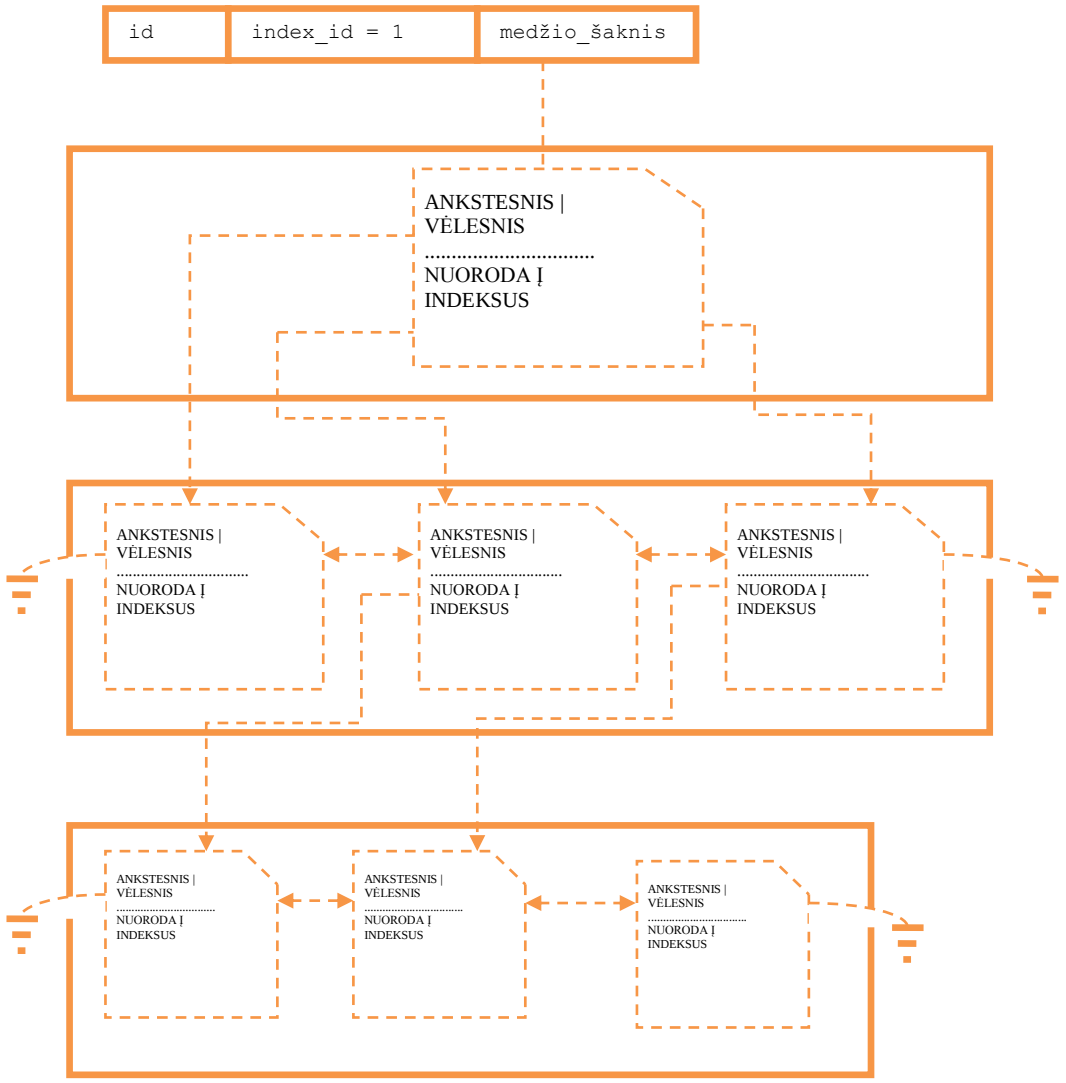


Pav. 5 Binarinis medis, suskaidytas į puslapius (Baniulis, Bajerio medis)

Pav. 5 parodytas binarinis medis, suskaidytas į puslapius; kiekvienas puslapis turi po 7 viršūnes. Todėl kiekvienas kreipinys į puslapį reikalauja vieno kreipinio į atmintį. Todėl žymiai sumažėja bendras paieškos laikas. Tarkim, puslapis turi 100 elementų, tada esant medyje 10^6 elementų, paieškai reikės atlikti vidutiniškai $\log_{100}(10^6) \approx 3$ (binarinio medžio atveju reikėtų $\log_2(10^6) \approx 20$ kreipinių į puslapius.

Microsoft SQL Server programinė įranga duomenis duomenų bazėje indeksuoja pasitelkdama B-medžių struktūrą (Microsoft). Kiekvienas B-medžio lapas (Baniulis, Ds2005/Teorija), (Baniulis, Bajerio medis) vadinamas indekso mazgu. B-medžio viršūnė (schemoje) vadinama medžio šaknimi. Žemiausio lygio mazgai vadinami lapais. Visi likę mazgai vadinami tarpiniais.

Taigi, (Baniulis, Bajerio medis) parodyti principai pritaikyti Microsoft SQL Server 2005 (taip pat ir keliose ankstesnėse versijose), žr. Pav. 6.



Pav. 6 Microsoft SQL Server duomenų indeksavimo principinė schema

Akivaizdu, kad mažesnis užklausų kiekis reikiamiems duomenims pasiekti yra tiesiogiai susijęs su laiko, tiems duomenims pasiekti, sąnaudomis. Taigi, Microsoft SQL Server programinėje įrangoje duomenų indeksavimui naudojami binariniai medžiai ir jų modifikacijos padeda pasiekti geresnių duomenų nuskaitymo iš duomenų sandėlio rezultatų laiko prasme.

2.6.2. Kubo skaičiavimų paralelizavimas

Duomenų kubu yra laikoma duomenų struktūra, susintetinta sujungus visus galimus grupavimo pagal pateikiamus parametrus variantus (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers). Natūralu, kad grupavimo uždavinius galima išskirstyti atskiriems procesoriams. (Dehne, Eavis ir Hambrusch), (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers), (Chen, Dehne ir Eavis), (Harinarayan, Rajaraman ir Ullman), (Muto ir Kitsuregawa) kalbama apie kubo skaičiavimo paskirstymą atskiriems kompiuteriams, naudojant keitimosi žinutėmis sistemą.

Duomenų struktūrai, turinčiai k atributų, skaičiuojant duomenų kubą, reikia atlikti 2^k operacijų. Kubas kiekvieną k atributą traktuoja kaip vieną matmenį k-matėje erdvėje. Atributų reikšmių agregatas atitinka vieną kubo erdvės tašką. Agregavimo funkcijos skirstomos į tris tipus:

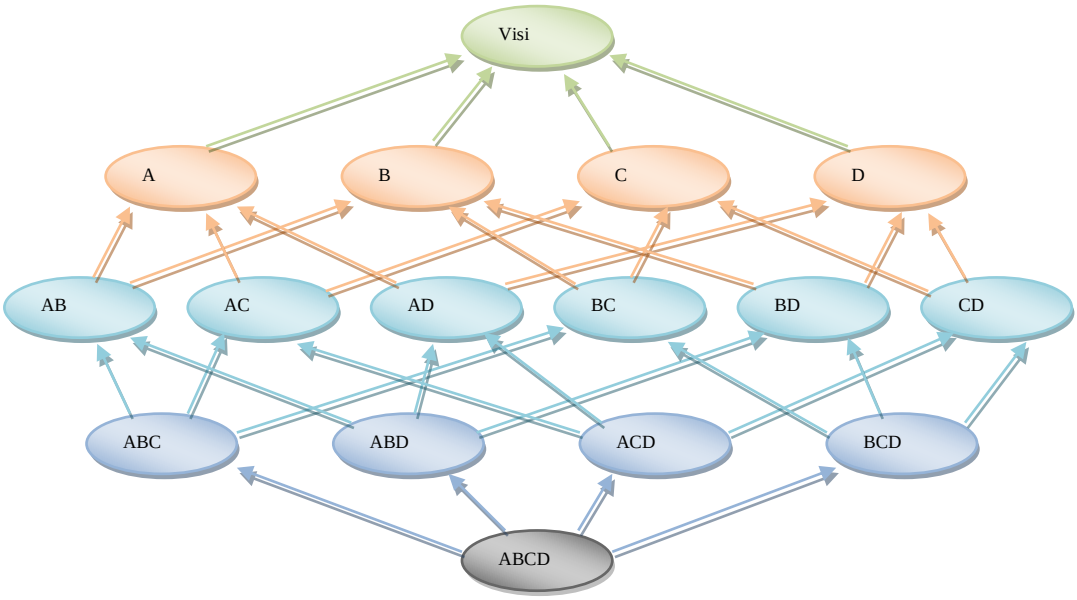
- perkeliamosios (*distributive*): jos gali būti suskaičiuotos atskirai, vėliau sujungtos. Tokios funkcijos yra suma, kiekis, minimumas, maksimumas.
- Algebrinės: jos gali būti suskaičiuotos iš perkeliamųjų funkcijų. Tokios funkcijos yra vidurkis, standartinis nuokrypis, n didžiausių reikšmių, n mažiausių reikšmių.
- Apibendrinančios (*holistic*): mediana, dažnis, rangas.

Procesorių tarpusavio komunikavimo problema apibrėžiama formule

$$T_{comm} = t_s + t_h d + t_w m \qquad \textbf{(4)}$$

kur t_s tai – pranešimo siuntimo / priėmimo inicijavimo (*handshake*) trukmė, t_h – tai dėl signalo keliavimo trukmės ir maršrutizavimo susidaręs vėlavimas, t_w – tai komunikacijos kanalo greitaveika, m – pranešimo dydis, d – tinklo mazgų skaičius.

Komunikavimo problemos (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers) siūlomos spręsti minimizuojant duomenų keitimąsi tarp procesorių. Duomenų keitimosi minimizavimui siūloma naudoti algoritmus, palyginančius duomenų rinkinių dydžius tarpusavyje ir nurodančius skaičiavimams naudoti mažesniuosius duomenų rinkinius, kaip pavaizduota Pav. 7.



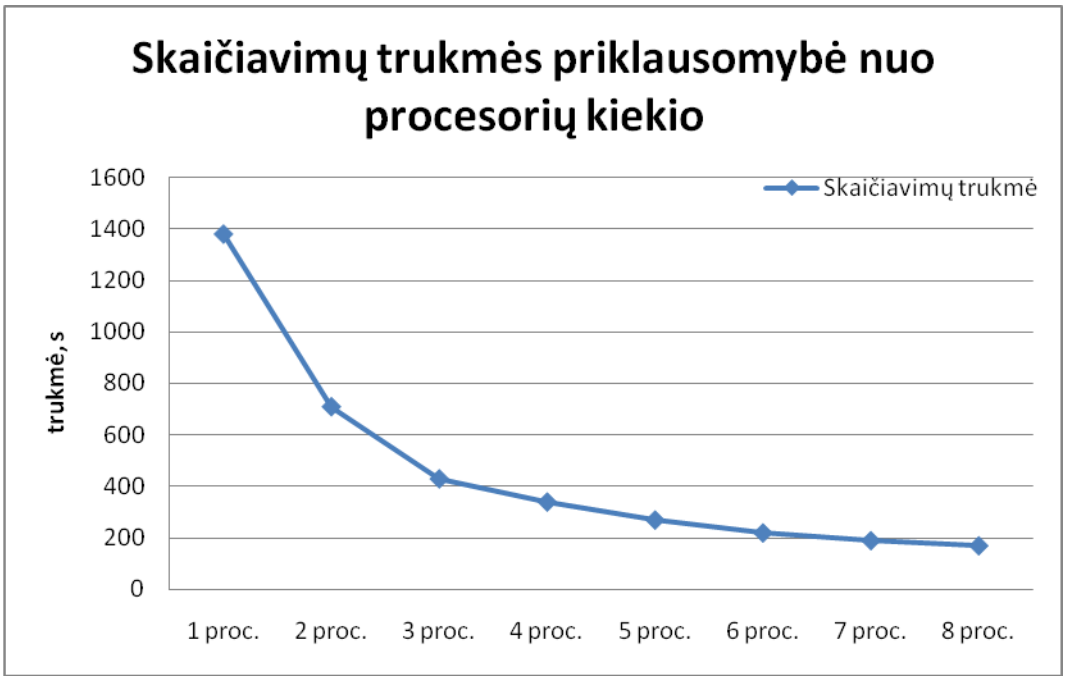
Pav. 7 Kubo grupavimo operacijų grafas

Kiekviena elipsė Pav. 7 atvaizduoja atitinkamą agregatą, o rodyklė nurodo galimą skaičiavimų kryptį. Kiekviename skaičiavimų lygyje (skirtingus lygius nusako skirtingos elipsių spalvos) procesoriai sprendžia kuri duomenų porcija yra mažiausia, kad iš jos būtų galima gauti aukštesniojo lygio rezultatus.

(Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers) skaičiavimų paralelizavimą skirsto į tokius uždavinius:

- duomenų padalinimas turimam procesorių kiekiui;
- duomenų persiuntimas procesoriams;
- duomenų agregavimas skirtinguose procesoriuose;
- konkrečių procesorių paskyrimas vykdyti konkrečioms užklausoms iš duomenų kubo;
- vietinių ir paskirstytų matmenų nustatymas valdančiame kompiuteryje.

(Dehne, Eavis ir Hambrusch) ir (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers) pateikia panašius kubų skaičiavimo trukmės sumažėjimo grafikus, tad žemiau pateikiamas apibendrintas skaičiavimų trukmės sumažinimo grafikas, skaičiuojant 6-ių matmenų duomenų kubą, sudarytą iš 1 000 000 įrašų.



Pav. 8 Paraleliniai skaičiavimai. Schematinė skaičiavimų trukmės priklausomybę nuo procesorių kiekio

Taigi naudojant papildomą techninę įrangą ir tam pritaikytą programinę įrangą galima sumažinti duomenų kubų konstravimo laiką (Dehne, Eavis ir Hambrusch), (Goil ir Choudhary, High Performance OLAP and Data Mining on Parallel Computers), (Goil ir Choudhary, A parallel scalable infrastructure for OLAP and data mining). Jei kubus skaičiuojanti techninė įranga naudojama ir kitiems tikslams, tai, atsižvelgiant į jos infrastruktūros apkrovimą, galima numatyti optimalų paros laiką kubo skaičiavimo pradžiai (Goil ir Choudhary, A parallel scalable infrastructure for OLAP and data mining). Kuomet techninė įranga yra mažiau apkrauta kitais darbais – skaičiavimai vyks greičiau.

2.6.3. Kubo dydžio mažinimas taip gerinant jo atsako laiko charakteristikas

Duomenų kubo dydis (tiksliau – užimama atmintis) yra vienas ir esminių parametrų, įtakojančių kubų skaičiavimo trukmę. Aukščiau minėti būdai siūlė taikyti įvairius kubo skaičiavimo metodus, tačiau dauguma jų reikalauja nemažų papildomų investicijų ir darbo laiko. Tačiau kai kuriais atvejais įmanoma sumažinti paties duomenų kubo apimtį ir taip paspartinti jo pirminio skaičiavimo laiką.

(Wang, Lu ir Feng) siūlo naują sąvoką – *koncentruotas kubas*. Koncentruotas kubas – tai pilnai suskaičiuotas kubas, kuris į vieną fizinį kortežą sutalpina semantiškai vienas kitą atitinkančių kortežų agreguotą rezultatą.

Tarkime turime lentelė *R*, kurioje yra tik vienas duomenų kortežas:

$$R(a_1, a_2, ..., a_n, m)$$

Tuomet *R* duomenų kubas bus sudarytas iš 2^n kortežų $(a_1, a_2, ..., a_n, V_1)$, $(a_1, *, ..., *, V_2)$, $(*, a_2, *, ..., *, V_3)$, ... $(*, *, ..., *, V_m)$, kur $m = 2^n$, * – visos stulpelio reikšmės. Kadangi *R* yra sudarytas tik iš vieno kortežo, tad $V_1 = V_2 = ... = V_m = arrg(r)$. Taigi fiziškai saugoti kubo duomenimis mums tereikia tik vieno kortežo $(a_1, a_2, ..., a_n, V)$, kur $V = aggr(r)$ aiškiai apibrėžiant, kad tai yra konkrečių kortežų grupavimo operacijos rezultatas.

Toks koncentruotas kubas išsiskiria sekančiomis savybėmis:

- *Koncentruotas kubas nėra suspaustas kubas*. Nors koncentruotas duomenų kubas dydžiu mažesnis už pilną duomenų kubą, tačiau jis nėra suspaustas. Taigi vykdant užklausas iš kubo nereikalingos papildomos operacijos duomenų išskleidimui.
- *Koncentruotas kubas – tai pilnai suskaičiuotas kubas*. Jis skiriasi nuo kai kurių autorių siūlomų dalinai suskaičiuotų kubų dydžio mažinimo strategijos, nurodančios pilnai suskaičiuoti tik dalį kubo kortežų.
- *Koncentruotame kube saugoma tiksli agreguota informacija*. Tuo jis skiriasi nuo kubų, kurių dydis sumažinamas naudojant aproksimaciją (Vitter, Wang ir Iyer), (Barbará ir Sullivan), (Shanmugasundaram, Fayyad ir Bradley) ir kt.
- *Koncentruotas kubas be pakeitimų gali būti taikomas su įprasta OLAP programine įranga.* Koncentruotame kube saugoma pilna reikiama informacija ir neapribojamas jos pasiekimas, kaip kad siūloma (Lee, Ling ir Li).

Grafinė (Wang, Lu ir Feng) siūlomo metodo atvaizdavimo forma pateikiama lentelėse žemiau.

Lentelė 1 Pavyzdinis duomenų rinkinys kubo konstravimui

ID	A	B	C	M
1	0	1	1	50
2	1	1	1	100
3	2	3	1	60
4	4	5	1	70

Lentelė 2 Pilnas kubas, sukonstruotas iš pavyzdinio duomenų rinkinio

ID	Kuboidas	A	B	C	M
1	Visos dimensijos	*	*	*	360
2	**ABC**	**0**	**1**	**1**	**50**
3	**A**	**0**	*****	*****	**50**
4	**AB**	**0**	**1**	*****	**50**
5	**AC**	**0**	*****	**1**	**50**
6	ABC	1	1	1	100
7	A	1	*	*	100
8	AB	1	1	*	100
9	AC	1	*	1	100
10	**ABC**	**2**	**3**	**1**	**60**
11	**A**	**2**	*****	*****	**60**
12	**AB**	**2**	**3**	*****	**60**
13	**AC**	**2**	*****	**1**	**60**
14	**B**	*****	**3**	*****	**60**
15	**BC**	*****	**3**	**1**	**60**
16	ABC	4	5	1	70
17	A	4	*	*	70
18	AB	4	5	*	70
19	AC	4	*	1	70
20	**ABC**	**6**	**5**	**2**	**80**
21	**A**	**6**	*****	*****	**80**
22	**AB**	**6**	**5**	*****	**80**
23	**AC**	**6**	*****	**2**	**80**
24	**C**	*****	*****	**2**	**80**
25	**BC**	*****	**5**	**2**	**80**
26	B	*	1	*	150
27	B	*	*	*	150
28	C	*	*	1	280
29	BC	*	1	1	150
30	BC	*	5	1	70

Lentelė 3 Koncentruotas duomenų kubas

ID	A	B	C	M	Kuboidų rinkinys
1	0	1	1	50	{A}, {AB}, {AC}, {ABC}
2	1	1	1	100	{A}, {AB}, {AC}, {ABC}
3	2	3	1	60	{A}, {AB}, {AC}, {ABC}
4	4	5	1	70	{A}, {AB}, {AC}, {ABC}
5	6	5	2	80	{A}, {AB}, {AC}, {ABC}
6	*	3	*	60	
7	*	3	1	60	
8	*	*	2	80	
9	*	5	2	80	
10	*	1	*	150	
11	*	5	*	150	
12	*	*	1	280	
13	*	1	1	150	
14	*	5	1	70	
15	*	*	*	360	

Iš Lentelė 2 ir Lentelė 3 bei jų tarpusavio ryšio matome, kad (Wang, Lu ir Feng) siūlomi kubų koncentravimo principai yra artimi Būlio algebroje taikomoms Būlio minimizavimo funkcijoms (Maciulevičius).

Duomenų išgavimas iš koncentruoto kubo atliekamas pasitelkiant paprastą operaciją (Wang, Lu ir Feng):

Koncentruoto kubo kortežo išplėtimo operacija kaip įėjimą naudodama nuorodą į konkretų kortežą ir nuorodą į kuboidą (arba jų rinkinį) sugeneruoja kortežų rinkinį, analogišką atitinkamame pilname kube esantįjį.

Taigi iš Lentelė 3 pateikto pavyzdžio kortežui, kurio *ID = 1* išplėtimo operaciją kuboidams *{A}, {AB}, {AC}* galima aprašyti taip:

išplėsti(kortežas(0, 1, 1, 50), {{A}, {AB}, {AC}}) =

Lentelė 4 Iš vieno koncentruoto kubo kortežo išplėstas kuboidas

Kuboidas	A	B	C	M
A	0	*	*	50
B	0	1	*	50
AC	0	*	1	50

2.7. Skyriaus išvados

Apžvelgus kelis galimus duomenų kubo skaičiavimo trukmės mažinimo būdus akivaizdu, kad kiekvienam jų įgyvendinti reikalingas papildomas darbas.

Didelių komercinių įmonių (Microsoft, Oracle) inžinierių suprojektuotų ir programuotojų sukurtų programinių produktų veikimo logikos tiesiogiai įtakoti, praktiškai, neįmanoma. Taigi, norint pasiekti užsibrėžtą tikslą (paspartinti duomenų kubų pirminius skaičiavimus) privalu naudotis duomenų šaltinio transformavimo metodais.

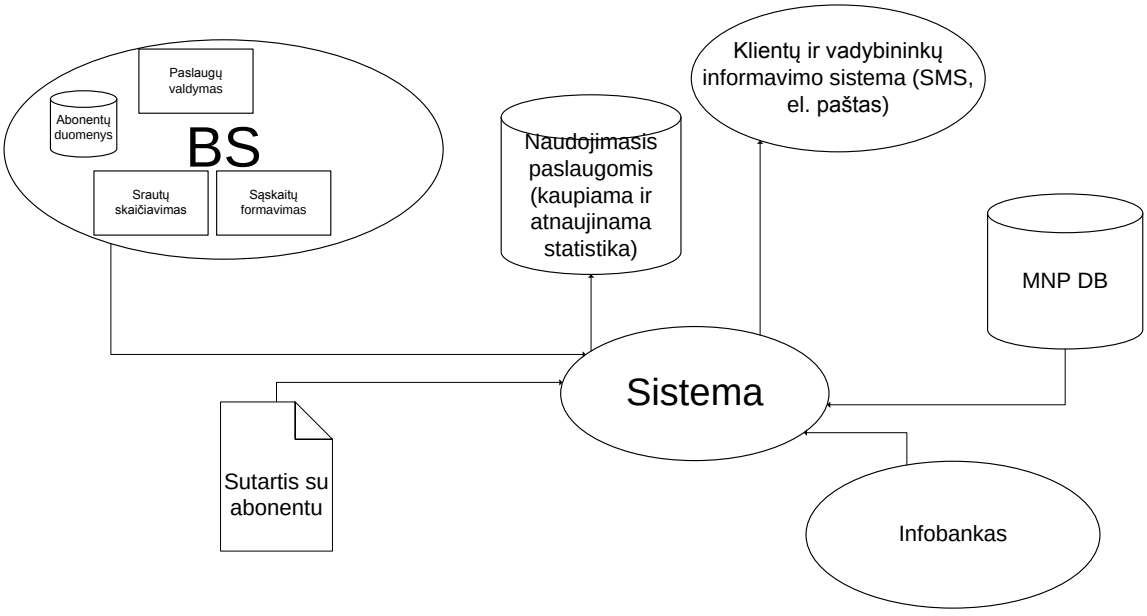
Taikant 2.5 skyriuje aptartus duomenų šaltinio transformavimo metodus galima su minimaliomis laiko sąnaudomis ir nedidele klaidų tikimybe pasiekti norimą rezultatą. O kurie iš 2.6 skyriuje pateiktų duomenų kubų skaičiavimo metodų panaudoti Microsoft SQL Server Analysis Services programinėje įrangoje, galima tik spėlioti. Greičiausiai minėtoje programinėje įrangoje įdiegti ir šie, ir kiti pažangūs duomenų kubų pirminio skaičiavimo optimizavimo metodai.

3. PROJEKTINIAI SPRENDIMAI

Šiame skyriuje aptariami esminiai programų sistemos kūrimo metu pritaikyti sprendimai, skirti sistemos suderinamumui ir optimaliai integracijai į egzistuojančią infrastruktūrą.

3.1. Sistemos kontekstas

Sukurtoji lojalumo administravimo sistema susieta su daugeliu įmonėje veikiančių verslo procesų. Sistemos aktualumas ir nauda verslui matuojamas konkrečiu darbo kiekio sumažėjimo administracijos darbuotojams ir vadybininkams.



Pav. 9 Sistemos kontekstas

BS – tai įmonės klientų valdymo ir apskaitos sistema. Abonentų naudojimosi paslaugomis statistikos skaičiavimui duomenys imami iš BS. Duomenų struktūrų vientisumą ir sistemų suderinamumą nuolatos užtikrinti yra gana sudėtinga dėl evoliucionuojančios BS sistemos. To siekiama sudarant šablonus duomenų vaizdavimo ir perteikimo formoms,

skirtoms betarpiškam sistemų bendravimui. Duomenų, reikalingų statistiniams lojalumo sistemoje skaičiavimams, apsikeitimas įgyvendinamas naudojant SSIS technologiją.

Į BS kuriamoji sistema teikia informaciją apie vadybininkų atliktus darbus ir abonentų pratęstas sutartis. Taip pat kuriamoji sistema atlieka reikalingas pelno / nuostolių analizes, kurių neatlieka turima BS.

Dėl vienos sistemos duomenų persipynimo su kita sistema ir neatsiejamo jų integralumo sukurtąją lojalumo sistemą galima laikyti BS posisteme, papildančia pastarosios galimybes ir įgalinančią paprasčiau bei laisviau planuoti lojalių abonentų išlaikymo sąnaudas.

Infobankas – tai didžiausia Lietuvoje duomenų bazė, teikianti informaciją apie fizinių ir juridinių asmenų skolas. Ši informacija lojalumų sistemai reikalinga ribojant ar bent jau perspėjant vadybininką apie galimai nemokų klientą. Naudojant operatyvų duomenų tarp BS, Lojalumo sistemos ir Infobanko keitimąsi įgyvendinamos prevencinės abonentų nemokumo priemonės.

MNP DB – tai centrinio Lietuvos viešojo ir viešojo judriojo telefoninio ryšio numerių tarptinklinių perkėlimų administratoriaus perkeltų numerių duomenų bazės kopija, pritaikyta BS ir Lojalumo sistemos reikmėms.

Būtent iš kitų mobiliojo ryšio operatorių, o ypač iš Omnitel ir Bitės migravę bei savo telefono numerio nepakeitę abonentai yra patraukliausia rinkos dalis. Šie vartotojai jau yra suformavę savo naudojimosi mobiliuoju telefonu įpročius ir jei jų netenkino buvusio operatoriaus teikiamų paslaugų kokybė ar kaina – vadinasi sprendimas nutraukti su juo sutartį buvo motyvuotas noru „kalbėti daugiau, mokėti mažiau“. Taigi remiantis padarytomis prielaidomis galima teigti, kad toks abonentas potencialiai pelningesnis, nei abonentas, kuris lig šiol nėra naudojęsis mobiliojo ryšio paslaugomis.

Skaičiuodama abonentų atneštą pelną Lojalumo sistema lygina mobiliojo numerio perkėlimo būdu tapusių abonentų ir naują numerį gavusių abonentų naudojimosi paslaugomis statistiką bei vėlesniuose konkretaus MNP atvejo metu skaičiuodama galimus pasiūlymus pratęsiant sutartį, taiko didesnį numanoma lojalumo koeficientą.

3.2. Sistemos mobilumo panaudojimui įgyvendinti keliami reikalavimai

Kliento kompiuteris gali būti bet koks kompiuteris ar kitas interneto naršymui skirtas įtaisas su modernią grafinę sąsają turinčia operacine sistemą, kurioje yra įdiegta moderni

žiniatinklio naršyklė. Naršyklė turėtų būti suderinama su JavaScript ir Cookies. Rekomenduojamos naršyklės yra: Microsoft Internet Explorer nuo 5.5 versijos, FireFox nuo 1.5 versijos, Opera nuo 8 versijos.

Taip pat norint spausdinti dokumentus kliento kompiuteryje turi būti įdiegta programinė įranga, galinti perskaityti *.PDF rinkmenas. Būtent PDF formatu bus pateikiami visi spausdintini dokumentai.

3.3. Pakankamai sistemos veikimo spartai už tikrinti keliami reikalavimai technikai

Šiame poskyryje trumpai aprašomi naudojamos techninės įrangos veikimo spartos parametrai.

3.3.1. Reikalavimai žiniatinklio serveriui

Žiniatinklio serveris yra suderinamas su Microsoft .Net Framework 2.0 darbo aplinka, taip pat jame įdiegta žiniatinklio serverio programinė įranga, Crystal Reports 9.0 versija. Pagal minėtos programinės įrangos gamintojų pateikiamus rekomendacinius reikalavimus aparatūrai, žiniatinklio serveris turi būti bent su 512 MB operatyviosios atminties, jo procesoriaus taktinis dažnis turi būti bent 800 MHz.

Naudojamas Intel Pentium 4 HT 2,8 GHz taktinio dažnio procesorių turinis kompiuteris su 2 GB operatyviosios atminties. Jame įdiegta Windows 2000 Server SBS operacinė sistema su IIS 6.0 žiniatinklio serverio programine įranga, Crystal Reports 9.0 versija. Taip pat kompiuteryje įdiegta Microsoft .Net Framework 2.0 darbo aplinka.

Naudojamas kompiuteris tenkina rekomendacinius žiniatinklio serveriui keliamus reikalavimus.

3.3.2. Reikalavimai duomenų bazės serveriui

Pagal planuojamus duomenų kiekius, skirtus apdorojimui (apie 50 mln. + 26 mln. įrašų apdorojami 7-niais – 4-riais pjūviais po du kartus per mėnesį su tendencija įrašų kiekiui augti po 15 procentų per ketvirtį metų) keliami tokie reikalavimai kompiuterio

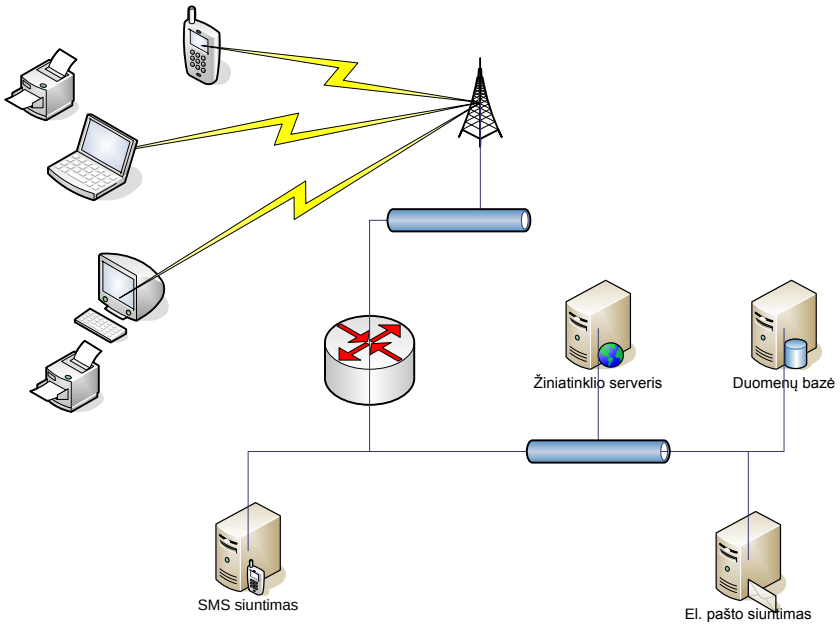
techninei įrangai: bent 3 GHz vieno branduolio Intel procesorius, bent 2 GB operatyviosios atminties, bent 100 GB laisvos vietos kompiuterio pastovioje atmintyje.

Duomenų bazės aptarnavimui naudojami du fiziniai kompiuteriai. Jų techninės ir programinės įrangos duomenys yra tokie: Intel Pentium Dual Core 2 Duo dviejų branduolių procesoriai, kurių vieno branduolio taktinis dažnis yra 2,4 GHz, 8 GB operatyviosios atminties, 100 GB - 1 TB laisvos vietos pastovioje kompiuterio atmintyje, Windows 2003 Server programinė įranga su Microsoft SQL Server 2005 DBVS.

Eksperimente taip pat papildomai naudojamas kompiuteris, kurio duomenys: Intel Pentium 4 2,667 GHz procesorius, 1 GB operatyviosios atminties, 80 GB - 120 GB laisvos vietos kompiuterio pastovioje atmintyje, Windows 2000 Server SBS operacinė sistema ir SQL Server 2000 DBVS.

3.4. Sistemos išdėstymo vaizdas

Stengiantis optimaliai išnaudoti sisteminius resursus programinė sistemos įranga buvo paskirstyta į keletą fizinių serverių, atsakingų už konkrečius programų sistemai keliamus reikalavimus. Pav. 10 pavaizduota sistemos išdėstymo schema. Galiniai vartotojų įrenginiai vaizduojami kaip stacionarūs, nešiojamieji ar delniniai kompiuteriai. Sutarčiai atspausdinti reikalingi spausdintuvai, tad schemoje pavaizduoti ir jie.

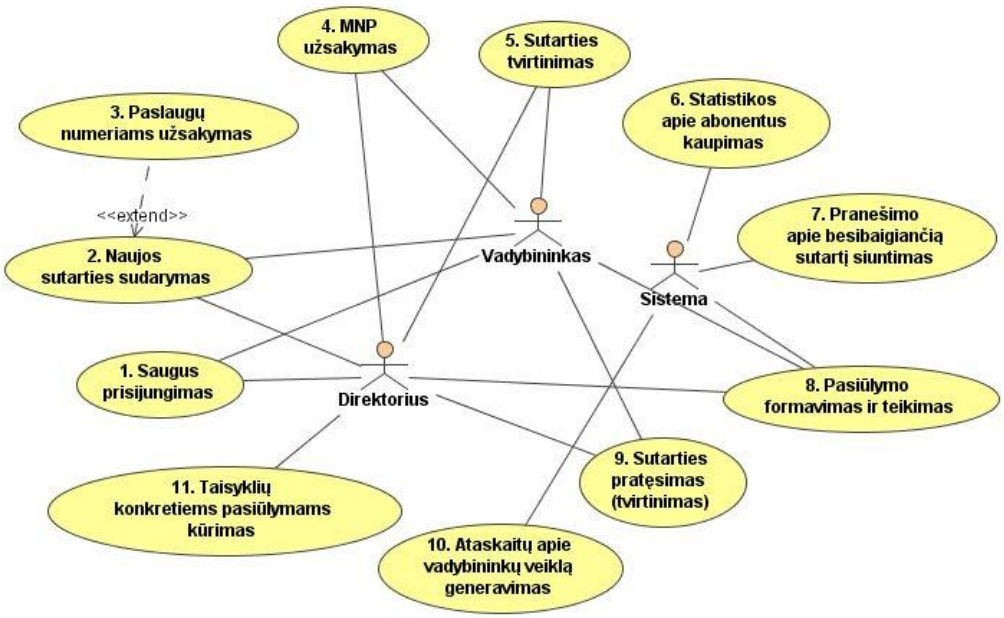


Pav. 10 Sistemos išdėstymo schema

3.5. Lojalumo sistemos panaudojimo atvejai

Lojalumo sistema naudojasi du vartotojų tipai. Tai įmonės vadovas (panaudojimo atvejų diagramoje – direktorius) ir pagrindinis sistemos vartotojas – vadybininkas. Vadybininkas panaudodamas jam sistemos suteikiamais įrankiais gali operatyviai atlikti visus reikiamus naujos sutarties sukūrimo veiksmus bei lojalaus abonento sutarties pratesimo veiksmus.

Vadybininkui aktualiausia informacija, pratęsiant lojalaus abonento sutartį, sistemos pateikiama kaip panaudojimo atvejų Nr. 6 bei Nr. 11 išeiga – panaudojimo atvejis Nr. 8, žr. Pav. 11. Panaudojimo atvejyje Nr. 8 pasiūlymo teikimas fiziškai įgyvendinamas patikrinant kokį lygį atitinka abonento naudojimosi mobiliojo ryšio operatoriaus paslaugomis naudingumo koeficientas. Tikrinama pagal direktoriaus sukurtą taisyklių rinkinį ir naudojimosi paslaugomis statistiką.



Pav. 11 Panaudojimo atvejų diagrama

Būtent panaudojimo atvejo Nr. 6 įgyvendinimas praktinėje projekto dalyje sukėlė daugiausiai problemų.

3.6. Dėmesys sistemos saugumui

Asmeninių duomenų konfidencialumas užtikrinamas vadybininko ir jo klientų lygiu. Kiekvienas vadybininkas, kuris sudarė sutartį su abonentu ar betarpiškai su juo bendrauja pratęsdamas sutartį gali peržiūrėti spausdinamą ar atspausdintą sutartį, kurioje yra asmeninių abonento duomenų. Tačiau ši veikla nenusižengia jokiems Europos Sąjungoje galiojantiems įstatymams ar teisės aktams, jei duomenys aptariami pokalbyje savanoriškai dalyvaujant abonentui.

Pagal Lietuvos respublikos įstatymus, įmonės, rinkdamos asmeninę informaciją apie savo klientus, privalo užtikrinti jų konfidencialumą, tad informacija pasinaudoti gali tik tie asmenys, kuriems ji yra skirta. Tą užtikrina įdiegta dviejų lygių sistema vartotojų autentifikavimo sistema.

Sistemos pakartotiniam panaudojimui pateikiami duomenys išlaiko integralumą. Jis vienareikšmiškai atitinka kliento įmonei suteiktus ir savo parašu ir / ar kitais atributais patvirtintus duomenis, kurie buvo pateikti pirmosios sutarties sudarymo metu ir / arba keisti nemokamai bet kada naudojimosi sutarties termino metu. Taip pat, klientui pageidaujant, pratęsiant naudojimosi paslaugomis sutartį, jo asmeniniai duomenys galu būti patikslinami ar papildomi be jokio papildomo mokesčio.

Duomenų pasiekiamumą teisėtiems vartotojams užtikrina sistemos principinė ir techninė veikimo logika. Esant būtinybei pagal Lietuvos Respublikos ar kitų ES ir / arba tarptautinių organizacijų teisėtus reikalavimus konkretūs duomenys pateikiami nesuteikiant minėtiems subjektams prieigos prie sistemos. Duomenys bus imami tiesiai iš duomenų bazės ir atrenkami pagal subjektų pateiktus kriterijus.

Sistema sukurta taip, kad net jos tiesioginiai naudotojai matytų tik reikiamus ir tik įstatymiškai leistinus naudoti duomenis. Kadangi priėjimas prie sistemos suteikiamas ne tik prie įmonės VPN tinklo su asmeniniais slaptažodžiais ir prisijungimo vardais prisijungusiems įmonės darbuotojams, tad besijungiantiems iš išorinio tinklo, privaloma antro lygio autentifikacija – vartotojai prie sistemos jungiasi suvedę į savo mobiliuosius telefonus gautus laikinus slaptažodžius.

VPN tinklas realizuotas naudojant MPLS technologiją per licencijuotus 5 GHz dažnio radijo tinklus ir licencijuotus 900 MHz / 1800 MHz GSM (CSD, HSCSD, GPRS, EDGE) bei 1900 / 2100 MHz UMTS tinklus. Duomenys taip pat apsaugoti 1024 bitų RSA

raktu bei 128 bitų SSL šifravimu. VPN tinklo naudojimas sumažina DoS tipo atakų grėsmę. Taip pat nuo internetinių įsilaužėlių VPN tinklą saugo Cisco maršrutizatoriai.

Vadybininkui pateikiami duomenys tik statistiniai abonento naudojimosi paslaugomis įverčiai ir paties vadybininko darbo planas ir minimali informacija, reikalinga abonento identifikavimui.

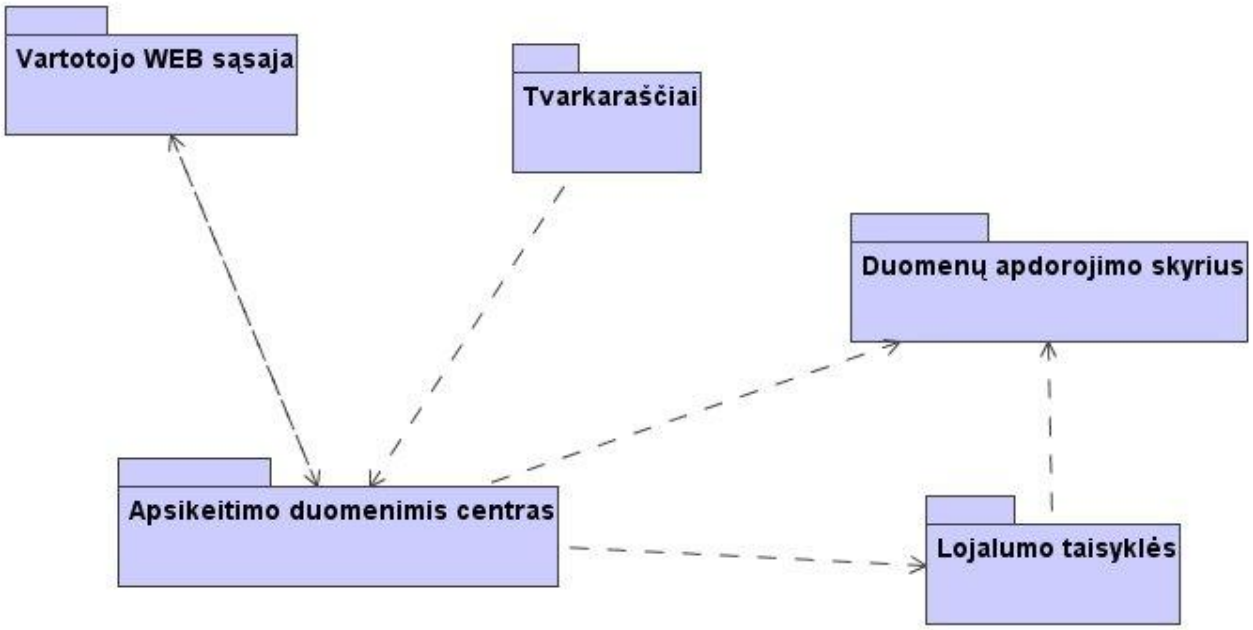
Taip pat galima teigti, kad duomenų perdavimas SMS pranešimu taip pat yra saugus jų pateikimo būdas dėl SMS siuntimui naudojamo SSL koduoto SMPP protokolo tiesioginio ryšio su SMSC.

3.7. Vartotojo sąsajos kalba

Kadangi sukurtas produktas naudojamas ne vien Lietuvoje, tad jame įgyvendintas vartotojo sąsajos suderinamumas su keliomis kalbomis. Ši priežastis įtakojo žiniatinklinės vartotojo sąsajos kūrimui pasirinkti ASP 2.0 technologija, kurią naudojant, nesunku kurti daugiakalbes sistemas, kurios pačios pagal vartotojo naršyklės nustatymus atpažįsta regioną, kuriame yra vartotojas ir pritaiko vartotojo sąsają tam konkrečiam regionui

3.8. Esminiai architektūriniai sprendimai

Šiame poskyryje pateikiama loginė sukurtos sistemos struktūra. Aprašomas sistemos suskaidymas į paketus ir juos sudarančias klases.



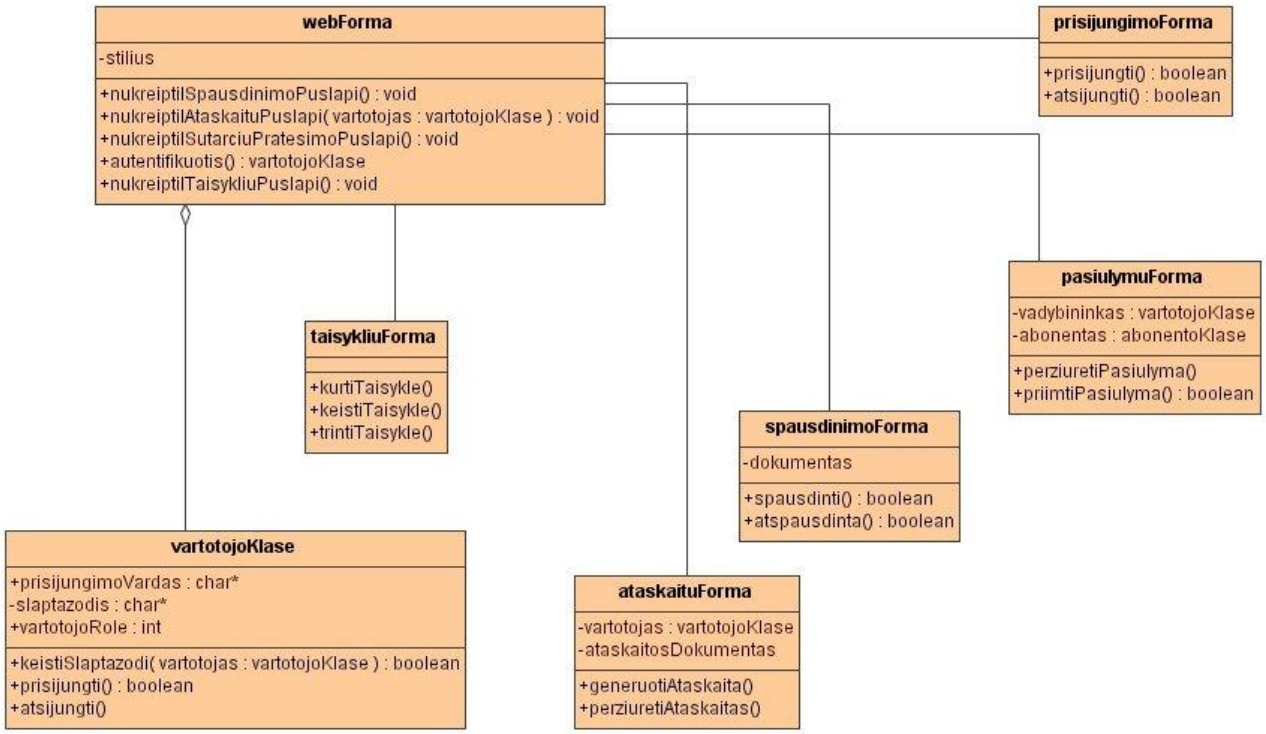
Pav. 12 **Sistemos išskaidymas į paketus**

Aukščiausiam hierarchiniam lygmenyje sistema išskaidyta į penkis paketus.

3.8.1. Paketų detalizavimas

Paketo **Vartotojo WEB sąsaja** klasių diagrama.

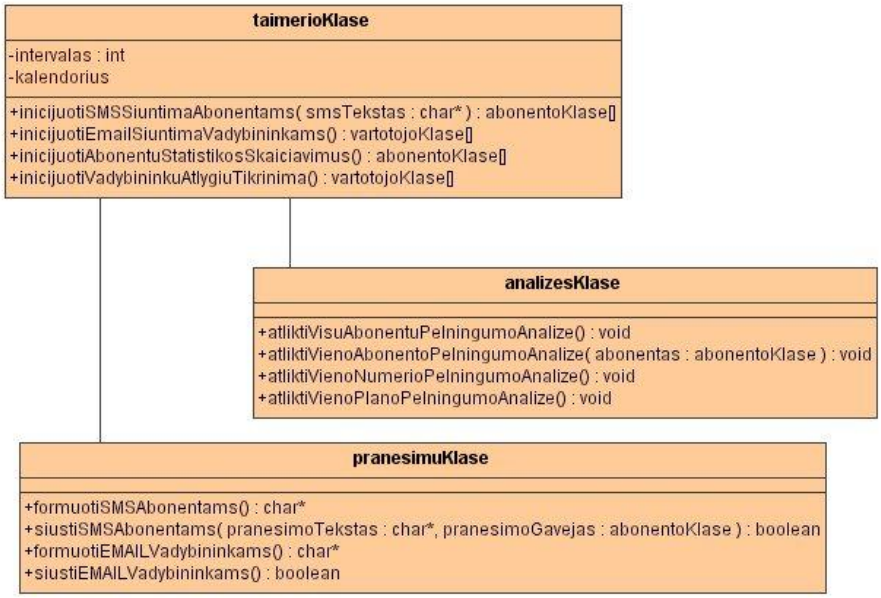
Pakete esančios klasės suteikia vartotojui galimybę prisijungti prie sistemos, naudotis jos teikiamais skaičiavimų rezultatais, atsispausdinti reikiamus dokumentus sutarties pratęsimui teisiškai įtvirtinti, vadybininkui peržvelgti savo atliktų darbų sąrašą, direktoriui peržvelgti visų vadybininkų atliktų darbų sąrašus, taip pat direktoriui formuoti arba keisti lojalumo nuolaidų taisykles.



Pav. 13 Paketo Vartotojo WEB sąsaja klasių diagrama

Paketo **Tvarkaraščiai** klasių diagrama.

Šiame pakete valdomos nuo laiko priklausančios funkcijos, tokios kaip pranešimų formavimo ir siuntimo vadybininkams bei abonentams sužadinimas, statistikos perskaičiavimo kiekvienam abonentui individualiai bei parenkant kitus perskaičiavimo kriterijus tam tikru mėnesio laiku.

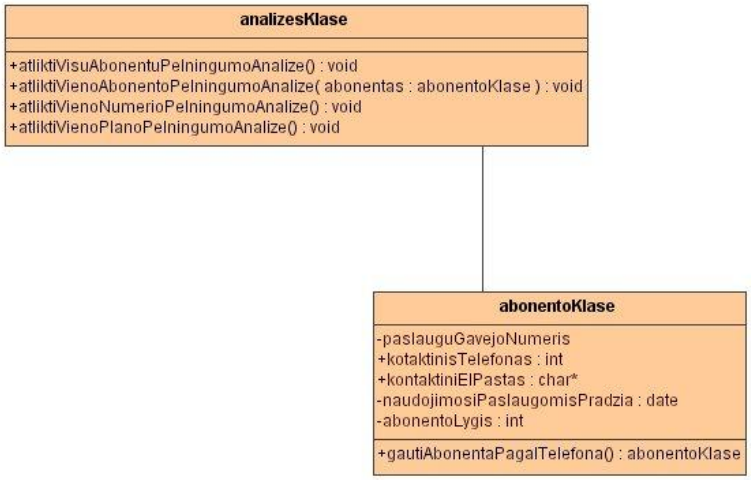


Pav. 14 Paketo Tvarkaraščiai klasių diagrama

Paketo **duomApdorojimo** klasių diagrama.

Pakete duomApdorojimo yra klasės, atsakingos už atliekamus skaičiavimus ir skaičiavimų rezultatus. Taip pat pakete duomApdorojimo yra klasė aprašanti individualų abonentą. Abonento objektas aprašomas su esminiais parametrais:

- paslaugų gavėjo numeris (abonento unikalus identifikatorius apskaitos sistemoje);
- kontaktinis telefono numeris (šiuo numeriu siunčiamas pranešimas apie besibaigiančią sutartį);
- kontaktinis el. pašto adresas (šiuo el. pašto adresu, jei jis nurodytas, siunčiamas priminimas su preliminariais pasiūlymais, skirtais konkrečiai šiam abonentui);
- naudojimosi paslaugomis pradžia (data, kada abonentas pasirašė pirmą sutartį su mobiliojo ryšio operatoriumi);
- abonento lygmuo lojalumo prasme (pagal šį parametrą sprendžiamas sutarties pratęsimo klausimas ir subsidijų, skiriamų abonentui dydis).

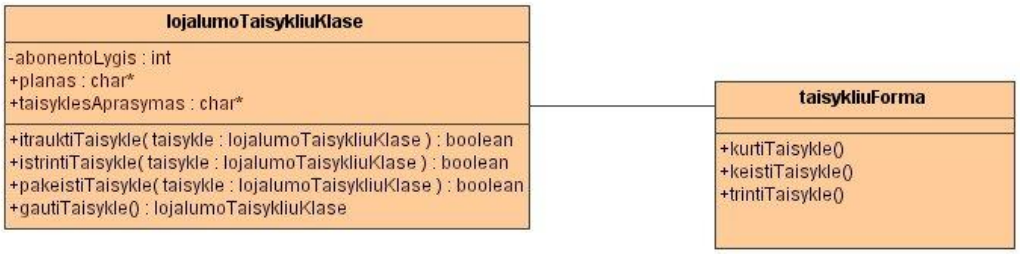


Pav. 15 Paketo duomApdorojimo klasių diagrama

Paketo **lTaisyklės** klasių diagrama.

Pakete lTaisyklės yra aprašomos funkcijos, skirtos lojalumo subsidijos dydžiui nustatyti konkrečiam abonentui. Taisyklės į sistemą yra įvedamos klasės lojalumoTaisykliuKlase pagalba. Šioje klasėje esantys metodai leidžia atlikti esamų taisyklių korekcijas taip pat, kaip ir šalinti sistemoje aprašytas taisykles (o tiksliau – padaryti jas nebegaliojančiomis), bei tokias pat taisykle kurti. Taisyklių kūrimui reikalingi duomenys yra abonento lygmuo ir abonento mokėjimo planas, kurio vartotojams bus taikoma atitinkama taisyklė. Taip pat numatyta vietoje mokėjimo plano keitimo suteikti abonentui galimybę pasiimti tam tikrą materialinę vertybę.

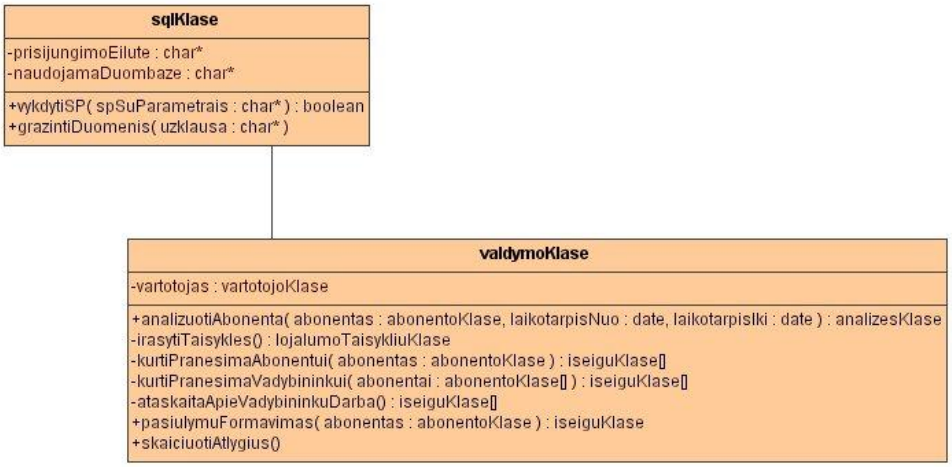
Kadangi aptarta veiksmų klasė glaudžiai susijusi su vartotojo įvedamais duomenimis, tad pakete lTaisyklės yra aprašoma ir vartotojo formos klasė, skirta taisyklių valdymui sistemoje.



Pav. 16 Paketo lTaisykles klasių diagrama

Paketo **valdymoCentras** klasių diagrama.

Pakete valdymoCentras esančios klasės yra atsakingos už tiesioginį ryšį su duomenų bazėmis ir atlieka tarpininko vaidmenį tarp vartotojo sąsajos, aukštesnės abstrakcijos klasių ir fizinio duomenų saugojimo lygmens.

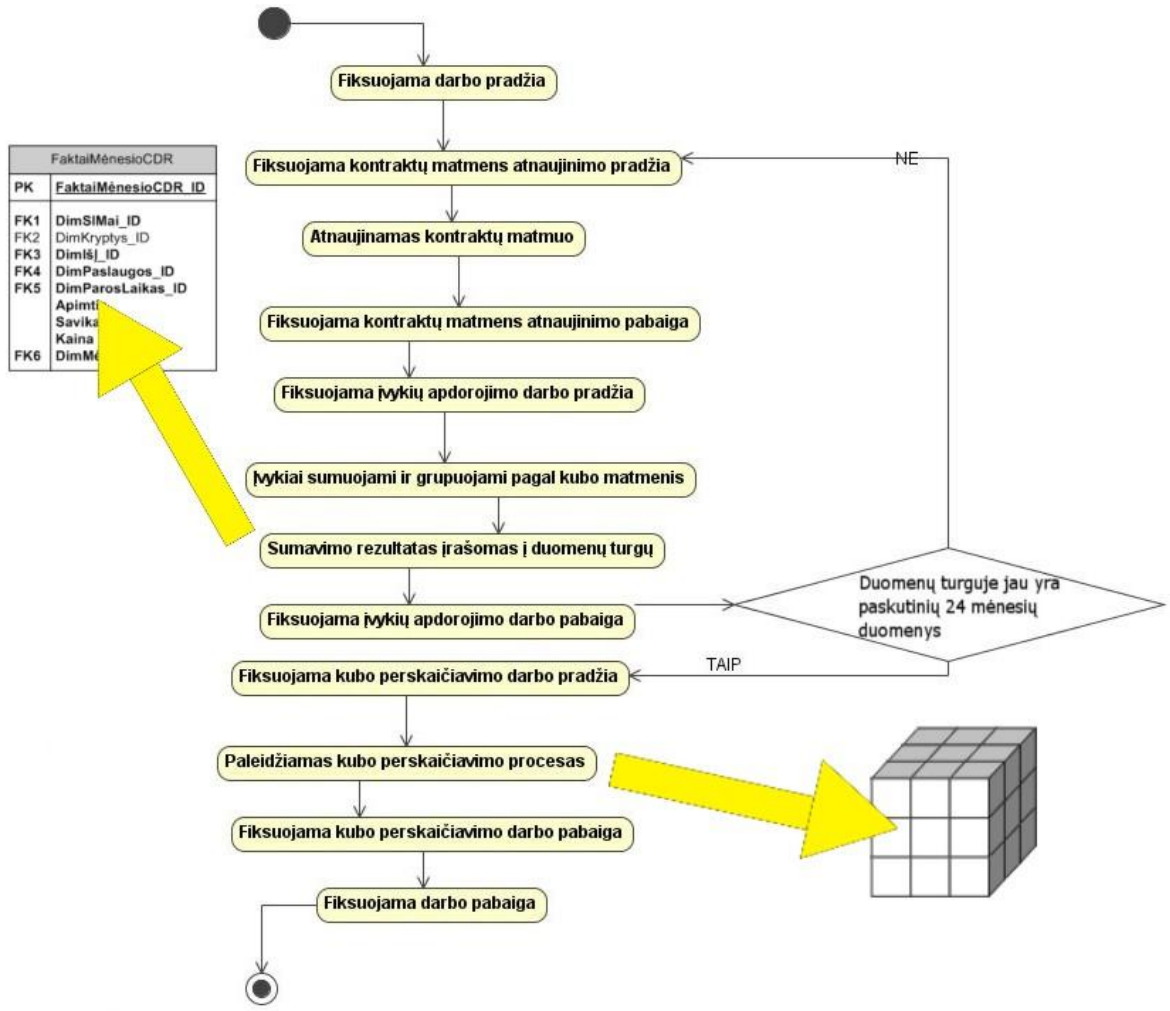


Pav. 17 Paketo valdymoCentras klasių diagrama

3.8.2. Skaičiavimų vykdymo veiklos diagrama

Skyriuje 4.1 (Eksperimento vykdymo schema) pateikiamo eksperimento veiksmų vykdymo veiklos diagrama (Pav. 18). Kadangi sukūrus sistemą duomenų turgus nebuvo užpildytas, tad privalu buvo atlikti 23 dalines ir vieną pilną iteracijas, žr. Pav. 18.

Veiksmų sekos vykdymas realizuotas SQL Server Integration Services (SSIS) užduočių seka (*task flow*). Vidiniai duomenų mainai kiekvienoje užduotyje įgyvendinami SSIS duomenų srautų seka (*data flow*).



Pav. 18 Statistikos kaupimo skaičiavimų vykdymo veiklos diagrama

Kiekvieną kartą kartojant iteraciją iš naujo parenkami seniausio (iš paskutinių 24) mėnesio duomenys, kurie dar nėra surašyti agreguota forma į duomenų turgų.

Jei dėl kokių nors priežasčių vieną mėnesį duomenų agregavimo procesas nebus sėkmingai atliktas, tad kitu tvarkaraščio paleidimo etapu iteracinis duomenų agregavimo procesas bus paleistas ne tik už naujausią mėnesį, bet ir už kitus, kurių duomenų reikia analizei.

„Paleidžiamas kubo perskaičiavimo procesas“ darbo punktas aktyvuoja SQL Server Analysis Services (SSAS) procesą. Šis, atfiltruodamas naujausių 24 mėnesių duomenis, į SSAS duomenų bazę vietoje senojo kubo įrašo atnaujintąjį.

4. DUOMENŲ ŠALTINIO REORGANIZAVIMO ĮTAKOS DUOMENŲ KUBO PIRMINIO SKAIČIAVIMO TRUKMEI EKSPERIMENTINIS TYRIMAS

Šiame skyriuje pateikiamas vykdytų eksperimentų aprašymas ir tarpusavyje palyginami jų metu gauti rezultatai. Pavyzdžiais pateikiamos rekomendacijos kaip galima sumažinti duomenų kubų konstravimo spartą.

Eksperimentai vykdomi kompiuteriu, kurio konfigūracija tokia: Intel Pentium Dual Core 2 Duo dviejų branduolių procesorius (taktinis dažnis yra 2,4 GHz), 8 GB operatyviosios atminties, 400 GB laisvos vietos pastovioje kompiuterio atmintyje, Windows 2003 Server operacinė sistema ir Microsoft SQL Server 2005 DBVS programinė įranga.

Pagrindiniams eksperimentams iš apskaitos sistemos pateikiami 24 mėnesių duomenys apie telekomunikacinius įvykius – iš viso 967 453 882 unikalių įrašų. Toks didelis įrašų kiekis yra visų aktyvių įmonės klientų (ne tik tų, kurių terminuota sutartis artėja į pabaigą). Visų abonentų naudojimosi paslaugomis duomenys reikalingi ir kitose analizėse, ir skirti ne tik lojalumo pasiūlymams formuoti, bet ir priimtų rinkodaros sprendimų pelningumo patikrai atlikti.

Pilnas dimensijų rinkinys – tai šešios dimensijos: *DimSIMai*, *DimKryptys*, *DimPaslaugos*, *DimParosLaikas*, *DimIšĮ*, *DimLaikas*. Sintezuotas dimensijų rinkinys – tai *DimSIMai* bei *DimMenesiai* dimensijos kartu su iš lėtai kintančių dimensijų sukonstruota sintetine dimensija DimSintetinė.

Minimalus analizei reikalingas kubas – tai minimalaus detalumo duomenų kubas, reikalingas sudaryti vieno abonento vieno mėnesio naudojimosi paslaugomis statistikai.

4.1. Eksperimento vykdymo schema

Buvo suruošta testavimo aplinka norint patikrinti 2.5.3, 2.5.4 skyriuose išsakytas hipotezes:

Hipotezė 1: *atlikus tarpinį duomenų agregavimą ir taip sumažinus duomenų šaltinio apimtį – sumažės ir duomenų kubo pirminio skaičiavimo trukmė***.**

***Hipotezė 2**: duomenų šaltinio struktūrą pakeitus kompaktiškesne dėl sumažėjusio skaitymo / rašymo (I/O) operacijų kiekio duomenų kubo pirminio skaičiavimo trukmė turėtų sumažėti.*

Testavimo proceso valdymas atliekamas Microsoft SQL Server Integration Services programinės įrangos. Testų vykdymo eiga:

1. Fiksuojama darbo pradžia
2. Fiksuojama kontraktų matmens atnaujinimo darbo pradžia
3. Atnaujinamas kontraktų matmuo
4. Fiksuojama kontraktų matmens atnaujinimo darbo pabaiga
5. Fiksuojama įvykių apdorojimo darbo pradžia
6. Iš duomenų sandėlio nuskaitomi įvykiai
7. Iš duomenų sandėlio nuskaitomos kubo dimensijos
8. Įvykiai sumuojami ir grupuojami pagal kubo matmenis
9. Sumavimo rezultatas įrašomas į duomenų turgų
10. Fiksuojama įvykių apdorojimo darbo pabaiga
11. Fiksuojama kubo perskaičiavimo darbo pradžia
12. Paleidžiamas kubo perskaičiavimo procesas
13. Fiksuojama kubo perskaičiavimo proceso pabaiga
14. Fiksuojama darbo pabaiga

Duomenų paruošimas duomenų kubo pirminiam skaičiavimui atliekamas 5-iais etapais:

- atnaujinamas kontraktų matmuo (3 testo žingsnis);
- iš duomenų sandėlio nuskaitomi įvykiai (6 testo žingsnis);
- iš duomenų sandėlio nuskaitomi kubo matmenys (7 testo žingsnis);
- įvykiai sumuojami ir grupuojami pagal kubo matmenis (8 testo žingsnis);
- sumavimo rezultatas įrašomas į duomenų turgų (9 testo žingsnis).

Nepriklausomai nuo to ar eksperimentas su dalininių duomenų agregavimu bei sintetine dimensija būtų vykdomas – 3 ir 6 žingsniai privalomi ir pilnam kubo skaičiavimui.

Žingsnyje **3. Atnaujinamas kontraktų matmuo** siunčiama užklausa į apskaitos sistemos duomenų bazę, duomenų sandėlyje dar neesantiems dimensijų įrašams gauti. Šios operacijos metu duomenų sandėlis susinchronizuojamas su realia klientų duomenų bazės

posisteme. Taigi duomenų sandėlis papildomas įrašais apie naujus klientus, o sutartis nutraukę klientai specialiai pažymimi, kad nebūtų įtraukinėjami į vėlesnes duomenų analizes.

Kadangi SQL Server Integration Services atliekami darbai apibrėžiami užduotimis (*task*), tad ir aprašant duomenų paruošimo (agregavimo) duomenų turgui procesą jį skaidome į kelias dalis. Taigi skirtingai nuo GROUP BY bei INNER JOIN naudojimo vienoje SQL užklausoje, įgyvendinant duomenų mainus tarp dviejų fiziškai ir logiškai atskirų serverių atliekami 6-9 eksperimento vykdymo schemoje paminėti veiksmai.

4.2. Eksperimento rezultatai

Šiame poskyryje pateikiami eksperimento rezultatai ir atliekamas kubo skaičiavimo trukmių palyginimas skaičiuojant pilną kubą su pilnu dimensijų rinkiniu, su sintetiniu dimensijų rinkiniu (žr. 2.5.4 skyrių, kuriame kalbama apie Duomenų turgaus faktų lentelės užimamos atminties minimizavimas) bei skaičiuojant minimalų analizei reikalingą kubą (žr. 2.5.2, Duomenų kubo konstravimas iš duomenų sandėlio, 2.5.3 Duomenų kubo konstravimas iš duomenų turgaus (data mart) skyrius) su pilnu dimensijų rinkiniu ir su sintetiniu dimensijų rinkiniu.

Laiko matavimo eksperimentų rezultatų failai pridedami priede Nr. 1 (kompaktinė plokštelė).

4.2.1. Kubo konstravimo trukmės priklausomybė nuo duomenų šaltinio dydžio

Pav. 19 Kubo konstravimo trukmės priklausomybė nuo duomenų šaltinio dydžio

Duomenų šaltinio su pilnu dimensijų rinkiniu dydį apribojus iki (apytikriai) 0,25 GB, 0,5 GB, 1 GB, 2 GB ir 4 GB buvo išmatuota kubų konstravimo trukmė. Eksperimentai parodė, kad kubo konstravimo trukmė logaritmiškai priklauso nuo kubo duomenų šaltinio dydžio.

Kadangi Microsoft SQL Server yra komercinis produktas, tad Microsoft neskelbia kokius kubo konstravimo trukmės optimizavimo būdus naudoja. Galima daryti prielaidą, kad Microsoft kombinuoja 2.6.1, 2.6.3 skyriuose paminėtus bei kitus galimus būdus.

4.2.2. Pilno kubo su pilnu dimensijų rinkiniu pirminio skaičiavimo trukmė

Vidutinė pilno duomenų kubo pirminio skaičiavimo trukmė yra **45 val. ir 39 minutės** (45,65 val.). Vidutinė trukmė apskaičiuota eksperimentą atlikus tris kartus, kas kartą matuojant jo vykdymo trukmę.

Kubo dydis, apskaičiuotas pagal formulę (1, yra: $n_{kubo} = 967453882 \times 732 \times 3 \times 5 \times 9 = 95\,603\,792\,619\,240$ unikalių reikšmių.

4.2.3. Minimalaus analizei reikalingo kubo su pilnu dimensijų rinkiniu skaičiavimo trukmė

Minimalus analizei reikalingas kubas konstruojamas laikantis šio dokumento poskyryje *2.5.3 Duomenų kubo konstravimas iš duomenų turgaus (data mart)* išsakytomis mintimis. Duomenų paruošimas duomenų turgui – tai vieno kalendorinio mėnesio įrašų apie telekomunikacinius įvykius agregavimo rezultato išsaugojimas.

Tiriamojo mobiliojo ryšio operatoriaus abonentai per mėnesį vidutiniškai sukuria apie 40 310 000 unikalių telekomunikacinių įvykių. Įvykius atspindinčius įrašus agreguojant aukščiau šiame poskyryje minėtu būdu gaunami 1 048 060 įrašai. Vieno mėnesio duomenų agregavimas trunka vidutiniškai **1 val. ir 32 min.** (1,53 val.).

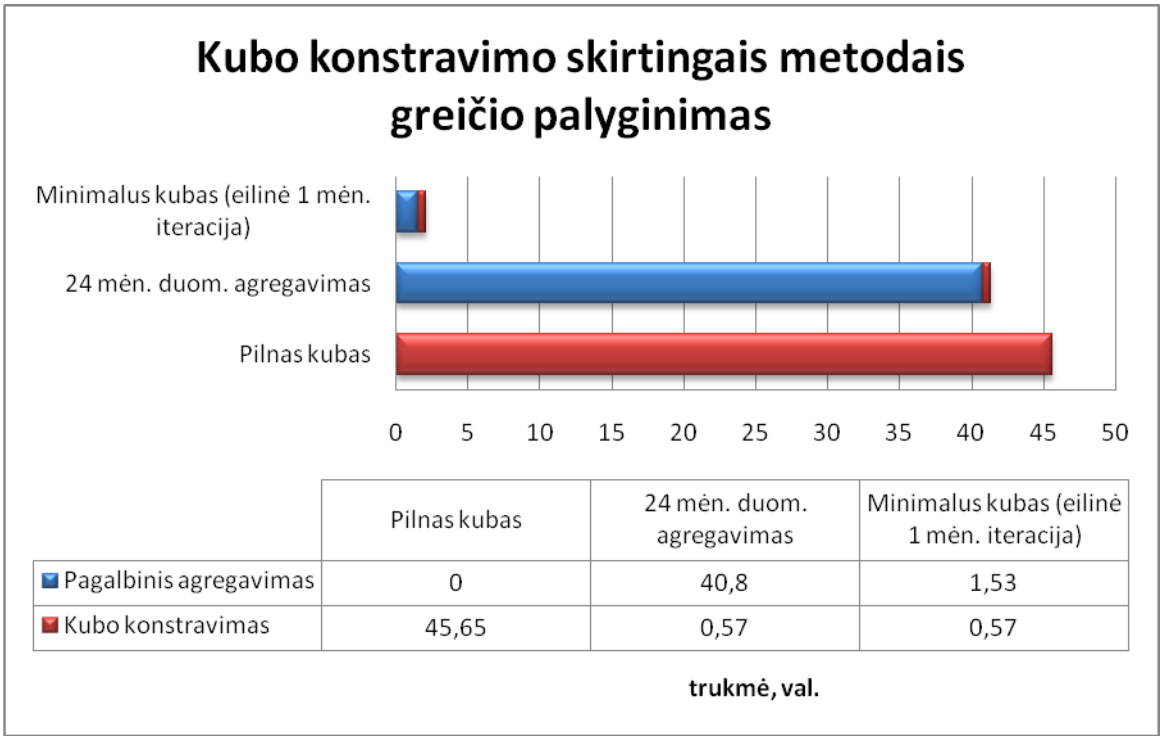
Tam, kad būtų galima atlikti eksperimentą reikėjo suagreguoti paskutinių 24 mėnesių duomenis. Nekreipiant dėmesio į žmogiškąjį faktorių ir padarytas pauzes tarp kiekvieno mėnesio duomenų perskaičiavimo proceso paleidimo bendra skaičiavimų trukmė yra apie **40 val. ir 48 min.** (40,8 val.), Pav. 20.

Kadangi aktualūs yra paskutinių 24 mėnesių duomenys, tad sekantį mėnesį reikės apskaičiuoti tik naujausio praėjusio mėnesio agregatus.

Bendras aktualių agreguotų faktų kiekis duomenų turguje yra 24 × 1 048 060 = 25 153 440 faktų. Tokio dydžio kubo pirminį skaičiavimą SQL Server 2005 įrankis SSIS testo konfigūracijoje paminėtame serveryje atlieka per **34 minutes** (0,57 val.).

4.2.4. Rezultatų, pateiktų skyriuose 4.2.2 ir 4.2.3 palyginimas.

Pastaba: visų kubų konstravimui naudojami paskutinių 24 mėnesių duomenys.



Pav. 20 Kubo pirminio skaičiavimo skirtingais metodais greičio palyginimas

Pilno kubo skaičiavimo trukmė yra didžiausia, nes jis konstruojamas iš neagreguotų faktų, esančių duomenų sandėlyje. Todėl konstruojamas duomenų kubas užima daugiausiai serverio atminties ir visas į ją netilpdamas operacinės sistemos yra rašomas į išorinę kompiuterio atmintį. Dėl didelio ilgai trunkančių kreipimųsi į išorinę atmintį kiekio kubo pilno pirminio skaičiavimo laikas yra didžiausias lyginant jį su kitais eksperimento metodais.

Duomenų turgus pildomas duomenimis 24-riomis iteracijomis. Po paskutinės iteracijos paleidžiamas minimalaus kubo pirminis skaičiavimas. Kadangi vieno mėnesio duomenų agregavimo operacija apdoroja, palyginus, nedidelį duomenų kiekį, tad dėl mažo įrašymo / skaitymo (*I/O*) operacijų kiekio ji vykdoma greitai.

Pabrėžtina, kad 24 duomenų turgaus paruošimo operacijas reikia atlikti vieną kartą. Vėliau duomenų turgus papildomas duomenimis už naujausią mėnesį. Vykdoma tik viena turgaus papildymo iteracija.

Kadangi duomenų turguje duomenų yra daug mažiau, tad minimalaus duomenų kubo konstravimas užtrunka trumpiau.

Eksperimento išeiga – tai nuo 95 603 792 619 240 iki 25 153 440 unikalių faktų sumažėjimas. Dėl to kubo pirminio skaičiavimo greitis nuo 45 val. 39 min. sutrumpėjo iki 34 minučių. Procentine išraiška – 1,24% pirminio vykdymo laiko. Tai 98,76% pagreitėjimas.

Įvertinus kiekvieną mėnesį reikiamą atlikti duomenų turgaus papildymo operaciją, kurios trukmė 1 val. ir 32 min., galutinis minimalaus kubo pirminio skaičiavimo trukmės sumažėjimas yra 95,4%.

4.2.5. Pilno kubo su maksimaliai sintetiniu dimensijų rinkiniu skaičiavimo trukmė

Naudojant 2.5.4 poskyryje pateiktą metodiką keturios dalykinės srities dimensijos buvo rankiniu būdu susintetintos į vieną. Pilno duomenų kubo su dirbtinai sumažintu užimamos atminties kiekiu pirminį skaičiavimą SQL Server programinė įranga atliko per, apytiksliai, **42 val. ir 17 min.** (42,28 val.). Tai vidutinė trijų eksperimentų vykdymo trukmė. Palyginus su 4.2.2 poskyryje pateikta pilno kubo su pilnu dimensijų rinkiniu skaičiavimo trukme, tai 7,4 % geresnis rezultatas.

4.2.6. Minimalaus kubo su maksimaliai sintetiniu dimensijų rinkiniu skaičiavimo trukmė

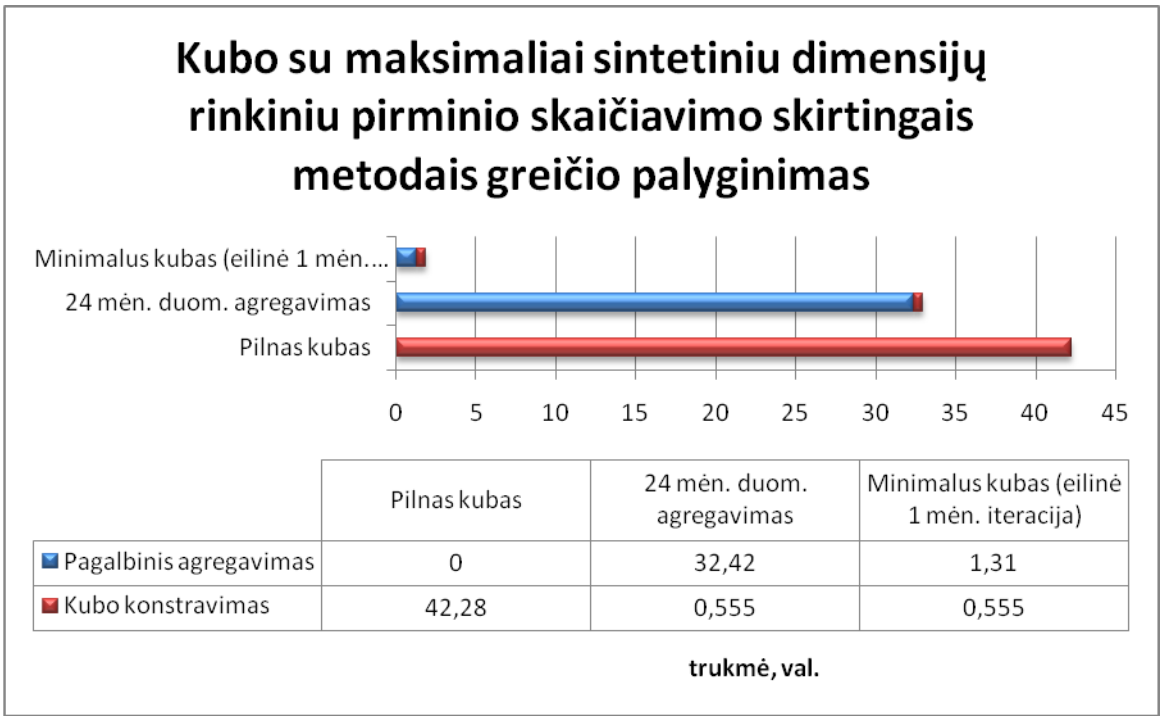
Duomenų turgaus (*data mart*) duomenų schema turi atitikti būsimą duomenų kubo struktūrą. Taigi, analogiškai 4.2.3 skyriuje aprašytam, duomenų turgaus už vieną mėnesį paruošimui naudojant sintetinę dimensiją vidutiniškai buvo sugaišta **1 val. ir 19 min.** (1,31 val.). Trukmė buvo gauta apskaičiavus trijų eksperimentų trukmių vidurkį. Tai 14,5% pagerėjimas, lyginant su analogišku eksperimento etapu, aprašytu 4.2.3 skyriuje.

Vadinasi, 24 mėnesių duomenys būtų agreguojami maždaug 24 kartus ilgiau. Panašų rezultatą pavyko pasiekti eksperimentu: **32 val. ir 25 min.** (32,42 val.). Tai 21,4% geresnis rezultatas, nei 4.2.3 skyriuje aprašytas analogiško eksperimento rezultatas.

Naudojant jau sukurtą sintetinę dimensiją (jos kūrimo metodika pateikiama 2.5.4 poskyryje) minimalaus reikalingo kubo (žr. 2.5.3 poskyrį) už vieną kalendorinį mėnesį pirminis skaičiavimas truko apie **33,3 min** (0,555 val.). Pateikiama vidutinė trijų bandymų trukmė. Palyginus su pirminio skaičiavimo, naudojant pilną dimensijų rinkinį, trukme (4.2.3 skyrius) nustatytas 2,6% skaičiavimų trukmės sumažėjimas.

4.2.7. Rezultatų, pateiktų skyriuose 4.2.5 ir 4.2.6 palyginimas.

Pastaba: visų kubų konstravimui naudojami paskutinių 24 mėnesių duomenys.



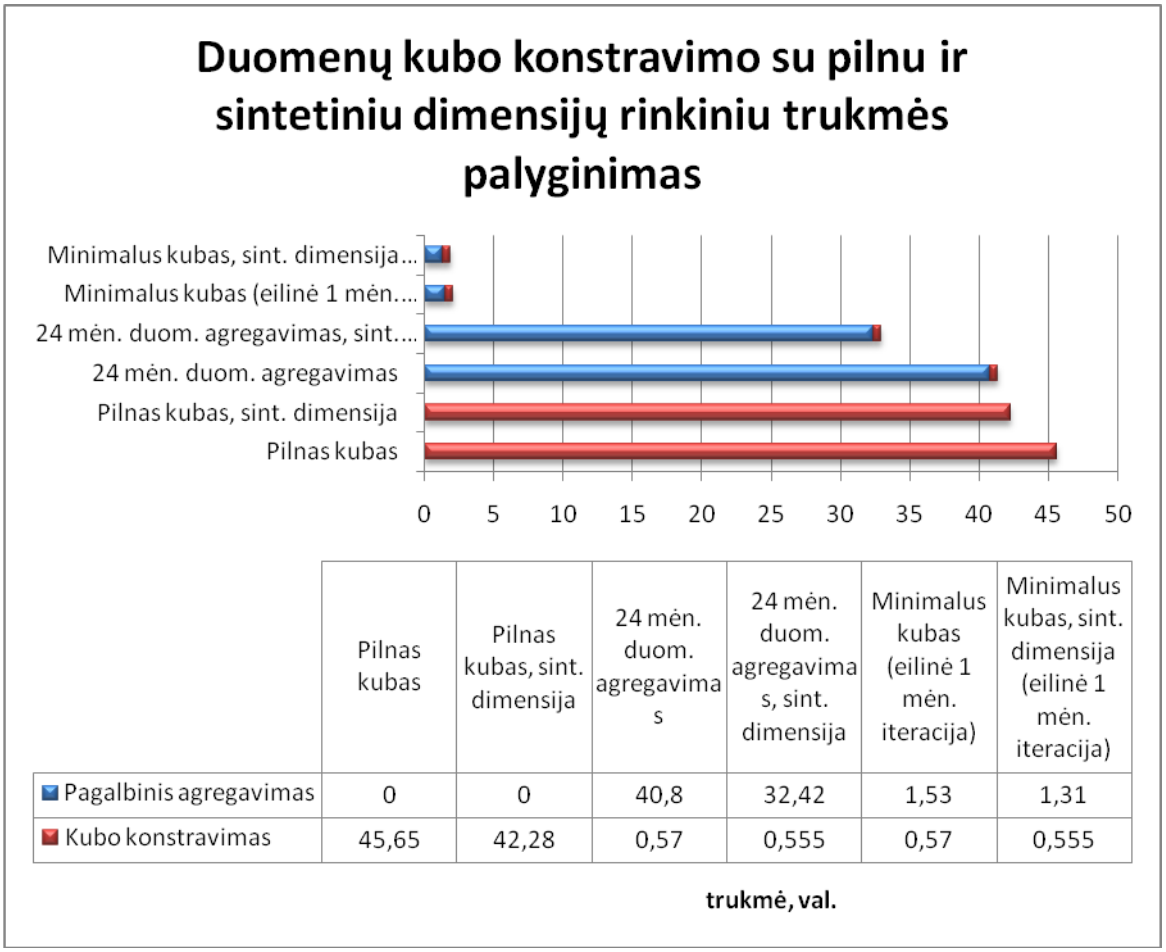
Pav. 21 Kubo su maksimaliai sintetiniu dimensijų rinkiniu pirminio skaičiavimo skirtingais metodais greičio palyginimas

Sumažinus kubo dimensijų kiekį (įvedus sintetinę dimensiją) dėl mažiau atminties užimančių pradinių duomenų pasiekiami geresni kubo pirminio skaičiavimo spartos rezultatai.

Geresni rezultatai buvo pasiekti ir duomenų perkėlimo iš apskaitos sistemos į duomenų turgų operacijos vykdymui, ir kubo konstravimui.

4.2.8. Kubo konstravimo trukmės su pilnu dimensijų rinkiniu ir sintetiniu dimensijų rinkiniu palyginimas

Pastaba: visų kubų konstravimui naudojami paskutinių 24 mėnesių duomenys.



	Pilnas kubas	Pilnas kubas, sint. dimensija	24 mėn. duom. agregavimas	24 mėn. duom. agregavimas, sint. dimensija	Minimalus kubas (eilinė 1 mėn. iteracija)	Minimalus kubas, sint. dimensija (eilinė 1 mėn. iteracija)
Pagalbinis agregavimas	0	0	40,8	32,42	1,53	1,31
Kubo konstravimas	45,65	42,28	0,57	0,555	0,57	0,555

Pav. 22 Kubo konstravimas su pilnu ir sintetiniu dimensijų rinkiniais

Iš palyginimo grafiko matome, kad sintetinės dimensijos sukūrimas naudingiausias vykdant žaliavinių duomenų įrašymą į duomenų turgų (24 mėnesių duomenų įrašymo trukmė sumažėjo 20,54%).

Tuo tarpu pilno duomenų kubo perskaičiavimo spartai sintetinės dimensijos turi mažesnę įtaką. Tačiau sintetinės dimensijos įvedimas ir šiuo atveju duoda apčiuopiamą naudą: skaičiavimų trukmė sumažėja 7,38%.

Minimalaus kubo pirminiame skaičiavime šio eksperimento metu žymiõs teigiamos įtakos sintetinės dimensijos įvedimas nepadarė. Eksperimentu nustatyta, kad minimalaus

kubo pirminio skaičiavimo trukmė sumažėja 2,63% (faktine išraiška pasiektas 40 sekundžių trukmės sumažėjimas).

4.2.9. Eksperimentų rezultatų apibendrinimas

1. Eksperimentais parodyta logaritminė priklausomybė tarp duomenų, iš kurių konstruojamas duomenų kubas, struktūros bei kiekio ir duomenų kubo konstravimo trukmės (Pav. 19).
2. Eksperimentų rezultatai rodo, kad naudojant sintetinių dimensijų kūrimo metodiką galima sutrumpinti duomenų turgaus paruošimo analizei (konstravimo) laiką. Eksperimente buvo pasiektas 7,4% efektyvumo padidėjimas konstruojant duomenų kubą iš ~95 GB duomenų bei 2,6% konstruojant kubą iš ~3,8 GB duomenų.
3. Iš eksperimentų aišku, kad sintetinės dimensijos įvedimo nauda tuo didesnė, kuo didesni duomenų kubai yra konstruojami. Šiuose eksperimentuose ~3,8 GB dydžio duomenims paversti kubu naudojant sintetinę dimensiją buvo sutaupytos 40 sekundžių; ~95 GB dydžio duomenis – sutaupytos 3,37 val.

5. IŠVADOS

1. Pasiūlyta sumažinti duomenų kubo konstravimui naudojamų dimensijų kiekį sujungiant lėtai kintančias dimensijas ir iš jų sukuriant naują sintetinę dimensiją.
2. Ištirta sintetinės dimensijos įvedimo įtaka duomenų kubo pirminio skaičiavimo trukmei ir parodyta, kad reikšminga teigiama sintetinės dimensijos įtaka pasireiškia konstruojant ypač didelius duomenų kubus ir agreguojant duomenis duomenų turgui.
3. Sukurtas praktinis pagrindimas automatinio sintetinių dimensijų generavimo įrankio kūrimui (žr. 2.5.4 skyrių, kuriame pateikiamas sintetinės dimensijos apibrėžimas bei patarimai automatinio įrankio kūrimui ir eksperimento rezultatų (4.2.9) 3-čias punktas).

6. SANTRUMPŲ IR TERMINŲ ŽODYNAS

SMS – „trumpinys nuo angliškų žodžių „Short Message Service“ (liet. „trumpųjų žinučių paslauga“), reiškiantis daugelyje mobiliųjų telefonų esančią paslaugą, skirtą trumpųjų žinučių siuntimui tarp mobiliųjų telefonų arba kitų panašių įrenginių. Trumpojoje žinutėje telpa 160 ženklų, tačiau kai kuriuose telefonuose jos gali būti ilgesnės, nes siųsdamas žinutę telefonas ją suskaido į keletą mažesnių“ [11].

SMSC – tai SMS centras – kompiuteris, kuris iš ryšio tinklo gavęs SMS pranešimą apskaičiuoja tolimesnę jo siuntimo ir pristatymo kryptį. Visi iš mobiliųjų telefonų ir kompiuterinių sąsajų išsiųsti SMS pranešimai keliauja į SMSC, jame saugomi ir persiunčiami galutiniam SMS pranešimo gavėjui. Jei gavėjo galinis įrenginys SMS pranešimo priimti tuo metu negali – SMSC saugo pranešimą atmintyje numatytą laiko tarpą kartkartėmis pabandydamas SMS pranešima išsiųsti. Jei po atitinkamo laiko tarpo (Lietuvoje visi operatoriai standartiškai naudoja 1 paros nustatymą) SMS pranešimas gavėjui nebuvo pristatytas – jis naikinamas SMSC įrenginyje.

BS – klientų duomenų valdymo ir apskaitos sistema. Angliškai – billing system. Tai įvairių rūšių paslaugų teikėjų naudojama apskaitos programų visuma, kurios teikia tikslią informaciją apie klientų naudojimąsi paslaugomis, jų apmokestinimą, sąskaitų klientams pristatymo vietas ir etc.

MNP – mobiliojo numerio perkėlimas, angliškai – mobile number portability. Tai techninis ryšio paslaugų teikėjų sprendimas, leidžiantis naudotis visomis teikiamomis paslaugomis pakeitus savo mobiliojo ryšio teikėją, tačiau nepakeitus savo mobiliojo ryšio telefono numerio.

VPN – „Virtualus privatus tinklas, tai tinklas, kuris naudoja žiniatinklį ar kitą viešą tinklo paslaugą kaip pagrindą sudaryti saugų susijungimą didelės erdvės tinkle“ rašoma (Vacca). VPN taip pat gali būti įrengtas naudojant bet kokius šiuolaikinius tinklų susijungimo metodus. VPN tikslas – saugiai pasiekti atitinkamus privačius resursus, tokius kaip įmonės vietinį kompiuterių tinklą.

MPLS – angliškai multiprotocol label switching. „Kompiuterių tinkluose ir telekomunikacijose MPLS yra duomenų nešėjo mechanizmas, kuris mėgdžioja kai kurias ryšio grandinėmis paremtų tinklų savybes naudodamasis paketinio duomenų perdavimo

tinklų privalumais. MPLS yra OSI modelio sluoksnis, kuris yra tarp tradicinio sluoksnio 2 (duomenų sąsajos sluoksnio) ir sluoksnio 3 (tinklo sąsajos sluoksnio) ir dažnai yra vadinamas „2,5 sluoksnio“ protokolu. MPLS buvo sukurtas teikti vieningą duomenų perdavimo paslaugą grandimis komunikuojamų tinklų ir paketais komutuojamų tinklų.“ (Davie ir Rekhter).

MHz – megahetrcai. Hz – hercas. Tai dažnio matavimo vienetas. Jis nurodo kiek kartų per sekunde atliekamas pilnas taktas. Pavadinta vokiečio fiziko Heinricho Rudolfp Hertzo garbei. Megahercas – tai milijonas hercų.

GSM – angliškai global system for mobile communications. Tai populiariausias mobiliojo ryšio standartas pasaulyje.

CSD – angliškai circuit switched data. Tai technologija, įgalinanti naudojantis mobiliuoju telefonu perduoti duomenis 9,6 kbps greičiu ir taip prisijungti prie žiniatinklio ar tk. Vartotojo įrenginiui susijungus su tinklo įrenginiais CSD kanalu jis yra rezervuojamas. Taigi CSD vartojimas dažniausiai apmokestinamas už susijungimo minutes.

HSCSD – angliškai high speed CSD. Tai CSD duomenų perdavimo technologija su patobulintu duomenų suspaudimo algoritmu taip pat įgalinanti įrenginį naudoti keletą komutuojamų susijungimų su tinklu ir taip keletą kartų padidinti duomenų perdavimo spartą. Priklausomai nuo mobiliojo ryšio operatoriaus leidimo naudotis ryšio kanalais HSCSD sparta gali siekti 115 kbps.

GPRS – angliškai general packet radio service. Tai GSM tinkle teikiama duomenų perdavimo paslauga. Skirtingai nei CSD – GPRS naudoja paketinį duomenų perdavimą ir nerezervuoja susijungimui atskiro balso kanalo. Maksimalus teorinis duomenų kanalo pralaidumas yra 171,2 kbps, bet realiomis sąlygomis pasiekiamas maksimumas yra apie 80 kbps.

EDGE – angliškai enhanced data rates for GSM evolution (EDGE) arba enhanced GPRS (EGPRS). Tai GPRS technologija, naudojanti modernesnį duomenų suspaudimo bei klaidų apdorojimo algoritmą. Maksimalus EDGE duomenų pralaidumas yra 236,8 kbps.

UMTS – angliškai universal mobile telecomunications system. Tai trečios kartos mobilusis ryšys. Jo veikimo principai skiriasi nuo GSM ryšio tinklo. Teoriškai UMTS tinklas duomenis vienu ryšio kanalu gali perduoti iki 11 Mbps, tačiau realiai įdiegti tinklai teikia 384 kbps susijungimo greitį.

RSA – tai vienas iš pirmųjų pakankamai patikimų duomenų kodavimo algoritmų, kuris vis dar laikomas saugiu naudojant ilgus kodavimo raktus. RSA algoritmą galima

naudoti ne tik elektroniniams parašams, bet ir duomenims koduoti (jis tam pakankamai spartus).

SSL – angliškai secure sockets layer yra technologija, suteikianti galimybę saugiai ir sparčiai perduoti duomenis tokiems poreikiams, kaip žiniatinklio naršymas, elektroninis paštas ir etc.

DoS – angliškai denial of service. Tai internetinių įsibrovėlių kompiuterinė ataka, kurios tikslas taip apkrauti kompiuterio ar tinklo resursus, kad jais negalėtų pasinaudoti tie, kuriems tai priklauso.

ASP – angliškai active server pages. Tai žiniatinklio technologija, įgalinanti kurti dinamines prieigas prie informacijos, pateikiamos vartotojo naršyklėje.

PDF (Portable Document Format) – tai kompanijos Adobe sukurtas dokumentų saugojimo formatas, kuris užtikrinta neiškraipytą dokumentų pateikimą bet kokioje kompiuterinėje platformoje.

Operatyvioji atmintis – tai speciali atminties sritis kompiuterio techninėje įrangoje, kuri skirta sparčiam duomenų pasikeitimui tarp kompiuterio centrinio procesoriaus ir išorinės atminties.

HT (HyperThreading) – Intel korporacijos sukurta technologija, kuri įgalina vieno branduolio procesorių lygiagrečiai vykdyti du procesus, naudojant tą pačią procesoriaus spartinančią atmintinę.

Telekomunikacinis įvykis – tai mobiliojo ryšio operatoriaus užfiksuotas skambutis, SMS ar MMS pranešimas, duomenų perdavimo sesija ar kitas aktyvus (inicijuotas) ar pasyvus (sulauktas) įvykis telekomunikaciniame tinkle.

Telekomunikacinio įvykio terminavimas – tai skambučio arba SMS bei MMS pranešimų gavimas galutinio telekomunikacinio įvykio gavėjo naudojamame tinkle.

SIM – abonento identifikavimo modulis. Angl. Subscriber Identification Module. Tokia mikroschema, kuri dedama į GSM tinklo telefono aparatą ir unikaliai identifikuoja abonentą.

Kontraktas – GSM tinkle tai viena SIM kortelė su paslaugomis, mokėjimo planu ir nuoroda į konkretų abonentą.

Abonentas – tai klientas, kuriam gali priklausyti keletas kontraktų.

UVG – uždara vartotojų grupė. Dažniausiai tai vienam abonentui priklausančių kontraktų rinkinys. Skambučiams ar kitoms paslaugoms UVG ribose mobiliojo ryšio operatoriai dažnai taiko specialias nuolaidas.

MSISDN – tai mobiliojo ryšio telefono numeris su tarptautiniu šalies kodu. Angl. Mobile Systems International Subscriber Identity Number.

IMSI – SIM kortelės serijos numeris. Angl. International Mobile Station Identifier.

OLAP – Analitinės sistemos. Angl. On-Line Analytical Processing.

OLTP – operacinių duomenų transakcijų apdorojimo. Angl. On-line Transaction Processing.

ETL – duomenų gavybos duomenų sandėliui ar duomenų turgui metodika, vadinama Išgauk / Transformuok / Įdėk. Angl. Extract / Transform / Load.

SSIS – SQL Server Integration Services. SQL serverio posistemė, skirta įvairių procesų automatizavimui ir tarpusavio sujungimui.

Lėtai kintanti dimensija – tai duomenų sandėlio terminologijos terminas, reiškiantis, kad naujų reikšmių atsiradimas dimensijoje yra mažai tikėtinas arba vyksta retai. Lėtai kintančios dimensijos pavyzdys gali būti: *lytis {moteris / vyras}*.

Kuboidas – duomenų sandėlio terminologijos terminas, reiškiantis duomenų kubo dalį, kuri apskaičiuota vykdant GROUP-BY operaciją tik pagal dalį dimensijų.

7. NAUDOTA LITERATŪRA

[1] **Agarwal, S. [et al.]** On the Computation of Multidimensional Aggregates. - Mumbai : Morgan Kaufmann, 1996. - ISBN:1-55860-382-4.

[2] **Baniulis, K.** Bajerio medis [Interaktyvus] // Duomenų struktūrų kursinis darbas. - KTU, 2006 m. balandis 29 d.. - [žiūrėta 2008-04-24]. - http://oras.if.ktu.lt/banikazy/ds/ds98r/DS/TEORIJA/t9bajer.html.

[3] **Baniulis, K.** Ds2005/Teorija [Interaktyvus] // oras.if.ktu.lt/banikazy/ds/. - KTU, 2006 m. balandis 29 d.. – [žiūrėta 2008-04-24]. - http://oras.if.ktu.lt/banikazy/ds/Ds2005/Teorija/DS_4_%20B_Medziai.ppt.

[4] **Barbará, D. ir M. Sullivan.** Quasi-cubes: exploiting approximations in multidimensional databases. - New York : ACM, 1997. - ISSN:0163-5808.

[5] **Chaudhuri, S. ir U. Dayal** An Overview of Data Warehousing and OLAP Technology. - New York : ACM, 1997. - ISSN:0163-5808.

[6] **Chen, Y.; Dehne F. ir Eavis T.** Parallel ROLAP Data Cube Construction on Shared-Nothing Multiprocessors. - Ansterdam : Kluwer Academic Publishers, 2004.

[7] **Dasu, T. [et al.]** Exploratory Data Mining and Data Cleaning . - San Francisco : John Wiley & Sons, 2003. - ISBN:0471458643.

[8] **Davie, B. ir Rekhter Y.** MPLS: technology and applications. - San Fransisco : Morgan Kaufmann, 2000. - ISBN:1-55860-656-4.

[9] **Dehne, F. [et al.]** Parallelizing the Data Cube. - London : Springer, 2001. - ISSN 0302-9743.

[10] **Goil, S. ir Choudhary A.** A parallel scalable infrastructure for OLAP and data mining. - Montreal : IEEE, 1999. - ISBN: 0-7695-0265-2.

[11] **Goil, S. ir Choudhary A.** High Performance OLAP and Data Mining on Parallel Computers. - Boston : Kluwer Academic Publishers, 1997.

[12] **Grey, J. [et al.]** Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals. - Redmond : Microsoft Research, 1995. - MSR-TR-95-22.

[13] **Harinarayan, V.; Rajaraman A. ir Ullman J.** Implementing data cubes efficiently [Žurnalas]. - New York : ACM, 1996 m.. - 2 : T. 25.

[14] **Kimball, R.** Slowly Changing Dimensions. - New York : DBMS Magazine, 2006.

[15] **Lee, S. Y.; Ling T. W. ir Li H. G.** Hierarchical compact cube for range-max queries. - San Francisco : Morgan Kaufmann Publishers Inc., 2000. - ISBN:1-55860-715-3.

[16] **Maciulevičius, S.** Kompiuterių teorija [Interaktyvus] // Stasio Maciulevičiaus namai. - KTU, 2005 m. sausis 27 d.. – [žiūrėta 2008-04-03]. - http://ifko.ktu.lt/~stama/Mokytojams/Komp_Teorija-BA.htm.

[17] **Microsoft** Clustered Index Structures [Interaktyvus] // SQL Server 2005 Books Online (September 2007). - Microsoft, 2007 m. lapkritis 01 d.. – [žiūrėta 2008-04-24]. - http://msdn.microsoft.com/en-us/library/ms177443.aspx.

[18] **Muto, S. ir Kitsuregawa M.** A Dynamic Load Balancing Strategy for Parallel Datacube Computation. - New York : ACM, 1999. - ISBN:1-58113-220-4.

[19] **Pedersen, T. B. ir Jensen C. S.** Multidimensional Database Technology. - Los Alamos : IEEE Computer Society Press, 2001. - ISSN:0018-9162.

[20] **Shanmugasundaram, J.; Fayyad U. M. ir S. Bradley. P.** Compressed Data Cubes for {OLAP} Aggregate Query Approximation on Continuous Dimensions. - New York : ACM, 1999. - ISBN:1-58113-143-7.

[21] **Vacca, J.** The Cabling Handbook (Second Edition). - Upper Saddle River : Prentice Hall PTR, 2001. - ISBN: 0-13-088317-4.

[22] **Vilimas, M.** Magistro darbas: Duomenų analizės priemonių tyrimas ir taikymas interneto sistemose. - Kaunas, 2004.

[23] **Vitter, J. S.; Wang M. ir Iyer B.** Data Cube Approximation and Histograms via Wavelets. - New York : ACM, 1998. - ISBN:1-58113-061-9.

[24] **Wang, W. [et al.]** Condensed Cube: An Effective Approach to Reducing Data Cube Size. - San Jose : IEEE publishing, 2002. - ISBN: 0-7695-1531-2.