

Andmeidad ja andmevakad

Teema 14

Argipäev

- ♦ Ettevõtetel võib olla palju andmebaase.
 - Nende struktuur on erinev.
 - Võivad olla realiseeritud erinevatel platvormidel.
 - Andmed võivad olla dubleeritud.
- ♦ Võibolla pole ettevõttel isegi terviklikku ülevaadet kõikidest enda käsutuses olnud infovaradest.
- ♦ **Kuidas selline olukord saab tekkida?**

Argipäev (2)

- ♦ Ettevõtted (eriti globaalsed) arenevad detsentraliseeritult – kohaliku haru juht peab kiiresti äri tööle saama.
- ♦ Kui peaksid kulutama aega enda andmebaaside struktuuri kooskõlastamisele tsentraalse skeemiga (nt tsentraalse põhiantmete andmebaasi struktuuriga), siis võtaks see palju aega.

Argipäev (3)

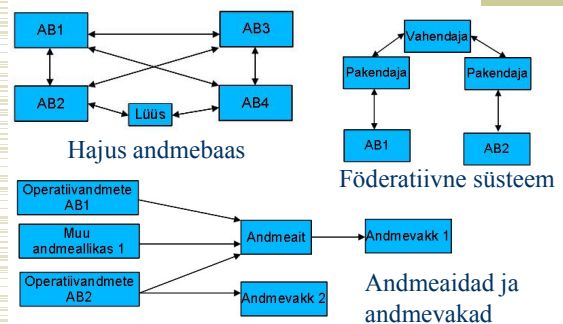
- ♦ Ettevõtted laienevad ka olemasolevate ettevõtete ülesostmise teel. Neil on oma IT süsteemid.
- ♦ Andmete koondamine, andmete struktuuri ühtlustamine/standardiseerimine on jällegi tagantjärgi tegevus.

Halb olukord

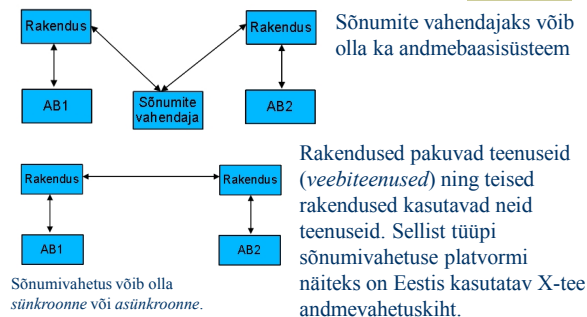


Informatsiooni saared

Kuidas saavutada andmete integratsioon?



Andmete integratsioon – sõnumivahetus rakenduste vahel



25.12.2017

Teema 14

7

X-tee

- ♦ Eestis väljatöötatud infosüsteemide andmevahetuskihit (edaspidi **X-tee**) on asutuste ja isikute vahelist turvalist ja tõestusväärtust tagavat internetipõhist andmevahetust ning riigi infosüsteemile turvalist juurdepääsu võimaldav tehniline infrastruktuur ja organisatsiooniline keskkond. (Vabariigi Valitsuse määrus *Infosüsteemide andmevahetuskihit*)

25.12.2017

Teema 14

8

Terminoloogia

- ♦ *Andmeait* (ingl *data warehouse*) asemel kohtab eestikeelses kirjasõnas ka mõisteid:
 - andmeladu,
 - andmevaramu,
 - andmehoidla.
- ♦ *Andmevaka* (ingl *data mart*) kohta on öeldud ka:
 - andmelett,
 - mini-andmevaramu,
 - andmekaubamaja.

25.12.2017

Teema 14

9

Operatiivandmete andmebaaside kasutamine

- ♦ Operatiivandmete andmebaase kasutavad süsteemid, mis on mõeldud **onlain-tehingutöötluks**:
 - tellimuste töötlemine,
 - arvete koostamine,
 - kauba laost väljastamine,
 - ...

25.12.2017

Teema 14

10

Operatiivandmete andmebaas

- ♦ Operatiivandmete andmebaas disainitakse, et ta võimaldaks kiiresti *otsida ja muuta üksikuid fakte (ridu SQL-andmebaasis)*.
- ♦ Organisatsioonil võib olla mitu operatiivandmete andmebaasi ja sinna võidakse koguda kattuvaid andmeid.
- ♦ Sellises andmebaasis on enamasti *aktuaalsed* andmed, mis annavad ülevaate organisatsiooni hetkeolukorrast, kuid mitte selle muutusest ajas.

25.12.2017

Teema 14

11

Operatiivandmete andmebaas (2)

- ♦ Otsustamine eeldab andmete *analüüsimist*.
- ♦ Mahukad ja keerulised päringud *koormavad* süsteemi.
- ♦ Operatiivandmete andmebaasi *füüsiline disain* pole selliste päringute jaoks optimeeritud.

25.12.2017

Teema 14

12

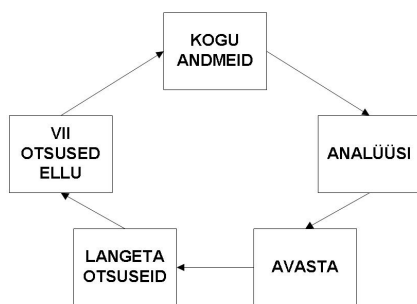
Näiteid otsuste langetajaid huvitavatest küsimustest

- ♦ Milliseid teisi tooteid ostetakse koos rukkileivaga?
- ♦ Millal ostab ostja kõige tõenäolisemalt piima?
- ♦ Milline on müügitulemuste kasv lõuna regioonis viimase 5 aasta jooksul võrreldes konkurendiga X?
- ♦ Mis on ühist meie parimatel klientidel?
- ♦ Kas mõni meie klientidest üritab tegeleda pettustega?
- ♦ Millise tõenäosusega täidame neljanda kvartali müügiprognoosi?

Tüüpilise päringu näide

- ♦ Onlain-tehingutöötluste süsteem:
 - `SELECT * FROM Arve WHERE arve_kood=2;`
- ♦ Andmete analüüsimiseks mõeldud süsteem:
 - `SELECT K.kauba_kood, K.kauba_artikkel, Count(*) AS arv FROM Kaup AS K INNER JOIN Arve_rida AS Ar ON K.kauba_kood=Ar.kauba_kood WHERE Year(Ar.loomise_aeg)=2011 AND Ar.summa>10000 GROUP BY K.kauba_kood, K.kauba_artikkel;`

Otsuste langetamise tsükkel



Analüüsiks ja otsustamiseks vajalikud andmed ...

- ♦ Erinevates andmebaasides.
- ♦ Erinevates arvutites.
- ♦ Halvasti ligipääseval kujul (erinevates failiformaatides).
- ♦ Võivad tulla (kusjuures väga suures mahus) uudsetest allikatest nagu *sotsiaalsõrgustikud*, *logid* ja *sensorid*.
- ♦ Andmebaasid on teostatud erinevate andmebaasisüsteemide abil.

Andmeait

- ♦ Andmeaidaks nimetatakse erinevatest operatiivandmete andmebaasidest (sealhulgas pärandsüsteemidest) ja välistest allikatest perioodiliselt uuendatavat *integreeritud* andmebaasi, mille alusel on võimalik teostada *juhtimisotsuseid*.
- ♦ Andmeaidas hoitakse nii ajaloolisi kui hetkel kehtivaid andmeid.

Välistest allikatest pärit andmed

- ♦ Andmed, mis on paberkandjal või elektroonisel kujul (nt veebiteenusena) kättesaadavad ajakirjandusest, teistest asutustest ja firmadest, Internetist või mujalt, ning mis on olulised juhtimisotsuste tegemisel.

Andmeait (2)

- ♦ Andmeait luuakse selleks, et pakkuda üleorganisatsioonilist **ühtlustatud** pilti kõigist andmetest.
- ♦ Andmeaidas on detailsed andmed (sama detailsusaste kui operatiivandmete andmebaasides) või mõõdukalt koondatud andmed.

Andmeait (3)

- ♦ Andmeaida all peetakse silmas teisest ehk sekundaarset andmekogumit, mis koondab esmastest (primaarsetest) andmekogudest saadud andmeid mingil kindlal eesmärgil. Andmeaita võib olla koondatud kas ühest või mitmest andmekogust pärit andmed.
- ♦ Andmeaidas võivad andmed olla **isikustatud** või **anonüümsed**.

Andmeaida omadused

- ♦ Arendamine on **pikaajaline protsess**.
- ♦ **Integreeritud** andmed erinevatest allikatest, et pakkuda tervikpilti organisatsiooni huvitavate olemite ning nendevaheliste seoste kohta.
- ♦ Suured **andmemahud**.
 - Detailsus * Ajaloolised andmed * Palju allikaid = Palju andmeid
 - Andmemahude kasv on väga kiire.
- ♦ Võib olla **tuletatud** andmeid – kontrollitud liiasus.

Andmeaida omadused (2)

- ♦ Eeldatakse, et lähteallikatest tulevad korrektsed andmed ja seetõttu ei pöörata erilist tähelepanu andmebaasi **kitsendustele**.
- ♦ Toimuvad **keerukad päringud**.
- ♦ Suhteliselt palju **indekseid**, et kiirendada päringuid.
- ♦ Andmeaita tuleb pidevalt andmeid **juurde laadida**, sest andmeallikateks olevad süsteemid "toodavad" kogu aeg üha uusi andmeid. Andmeaidas hoitakse nii aktuaalseid kui ka ajalooliseid andmeid.

Andmeaida omadused (3)

- ♦ Andmeid üldjuhul **ei kustutata** – andmebaasis ka ajaloolised andmed.
- ♦ Kustutamise otsust mõjutavad:
 - **ülisuured andmemahud** (andmetega ei osata midagi peale hakata, pole majanduslikult otstarbekas pikka aega säilitada),
 - **seadusandluse nõuded** (andmeid tuleb säilitada teatud ajaperioodi).

Operatsioonide sagedus andmeaidas

- ♦ Põhilised operatsioonid INSERT ja SELECT.
- ♦ DELETE harva.
- ♦ UPDATE peaaegu mitte kunagi.

Andmete säilitusperioodi näide – raamatupidamise seadus

- ♦ § 12. Raamatupidamise dokumentide säilitamise kohustus.
 - Raamatupidamisregistreid, mis on vajalikud majandustehingute arusaadavaks kirjeldamiseks revideerimise käigus, peab raamatupidamiskohustuslane säilitama **seitse aastat** alates vastava majandusaasta lõpust.
 - Raamatupidamisregistreid, mis on loodud elektrooniliselt, on raamatupidamiskohustuslane kohustatud ka säilitama elektrooniliselt. Elektrooniliste andmete loetavus peab olema tagatud kogu säilitusaja jooksul.

Andmete säilitusperioodi näide – isikuandmed

- ♦ Isikuandmete osas tuleneb säilitustähtaja kindlaks määramise nõue isikuandmete kaitse seaduse § 6 punktide 2 ja 3, mis sätestavad eesmärgikohasuse ja minimaalsuse põhimõtte.
- ♦ See tähendab minimaalsust ka ajalisel mõõtmel.
- ♦ Andmeid ei tohi töödelda kauem (säilitamine on andmete töötlemise üks viis!), kui see on vajalik andmete kogumisel seatud eesmärgi saavutamiseks.

Allikas: Andmekogude juhend. Andmekaitse inspeksioon. 14.08.2013.

Isikuandmete kaitse seadus

- ♦ § 16 (4) Kogutud isikuandmeid on lubatud töödelda teadusuuringu või riikliku statistika vajadusteks, olenemata sellest, millisel eesmärgil neid isikuandmeid algselt koguti. Teadusuuringu või riikliku statistika vajadusteks kogutud isikuandmeid on kodeeritud kujul lubatud säilitada ka hilisemate teadusuuringute või riikliku statistika vajadusteks.

Andmete säilitamine – vastuolulised jõud

- ♦ Säilita võimalikult palju andmeid võimalikult kaua, et andmetest võimalikult palju „välja pigistada“. Võib anda konkurentsieelise, suurendada kasumit, kasutajate rahulolu.
- ♦ Ära kogu andmeid, mida pole tööks vaja ja hoia kogutud andmeid vaid nii kaua kui vajalik eesmärkide täitmiseks, et vähendada eraelu puutumatuse riivet ja mitte pakkuda liiga detailset „suurt pilti“.

Mõõtühikud

- ♦ 1 terabait = 1024 gigabaiti
- ♦ 1 petabait = 1024 terabaiti
- ♦ Konsultatsioonifirma Gartner hinnangud andmeite suurusele.
 - Väike andmeait – vähem kui **5 terabaiti** andmeid
 - Keskmise suurusega andmeait – **5–20 terabaiti** andmeid
 - Suur andmeait – **üle 20 terabaidi** andmeid

Teradata Petabyte Power Players club (2008. aasta sügise seisuga)

- | | |
|---|---|
| <ul style="list-style-type: none"> ♦ eBay <i>5 petabaiti</i> andmeid ♦ Walmart <i>2.5 petabaiti</i> andmeid ♦ Bank of America <i>1.5 petabaiti</i> andmeid ♦ Nimetu finantsteenuseid pakkuv kompanii <i>1.4 petabaiti</i> andmeid ♦ Dell <i>1 petabait</i> andmeid | <ul style="list-style-type: none"> ♦ Teradata – andmebaasisüsteemide arendaja, mis spetsialiseerunud <i>väga suuri andmebaase</i> (andmemah, kasutajate hulk, päringute arv) toetavate andmebaasisüsteemide loomisele. |
|---|---|

eBay andmeidad (2009. aasta kevade seisuga)

- ♦ Andmeait 1
 - >2 petabaidi kasutajate **andmeid**
 - Kümned kuni tuhanded **kasutajad** iga päev
 - Miljonid **päringud** päevas
 - Andmed tulevad sadadest erinevast **lähteandmebaasist**
 - Andmeidast lähevad andmed kümnetesse **andmevakkadesse**, milles tavaliselt alla 5 terabaidi andmeid ning sageli isegi alla 500 gigabaidi andmeid
 - Põhineb *Teradata* andmebaasisüsteemil

eBay andmeidad (2009. aasta kevade seisuga) (2)

- ♦ Andmeait 2
 - Veebi ja arvutivõrgu sündmuste **logi**
 - Viimase 90–180 päeva kohta väga **detailed** andmed
 - 4.5 petabaiti **andmeid** (kasutatakse pakkimist)
 - Väike arv **paralleelseid kasutajaid**
 - Põhineb *Greenplum* andmeaitade halduse süsteemil (põhineb PostgreSQLil)
 - Edastab andmeid Teradatal põhinevasse andmeaita

Andmeaitade suuruselt 2017. aasta seisuga

- ♦ Suured andmeidad võivad sisaldada **rohkem kui 100 terabaiti** andmeid.
- ♦ Sageli on *vähem kui kolmandik* andmetest *tegelikud lähtesüsteemidest pärit andmed* – ülejäänud andmemaht tuleneb andmeida disaini spetsiifikast (näiteks on tabelitele loodud palju indekseid).

Guinnessi rekordite raamat – maailma suurim andmeait (2012)

- ♦ Rekord aastast 2012
- ♦ Andmeida andmemaht **3** petabaiti
- ♦ Andmebaasisüsteem IBM DB2
- ♦ Asukoht – Iirimaa
- ♦ Ilmselgelt leidub ka suuremaid kuid nende omanikud pole tahtnud rekordite raamatusse sattuda.

Guinnessi rekordite raamat – maailma suurim andmeait (2017)

- ♦ Rekord aastast 2014
- ♦ Andmeida andmemaht **12.1** petabaiti
- ♦ Andmebaasisüsteem: SAP-HANA
 - Tabeli esmakordsel lugemisel loetakse kogu tabel muutmällu (*in-memory system*)
- ♦ Koostöö: SAP, BMMsoft, HP, Intel, NetApp ja Red Hat
- ♦ Asukoht – USA

Maailma suurim andmeait (2017) (2)

- ♦ Laborieksperiment, mitte reaalsed andmed.
- ♦ Riistvara kaalus kokku ligi kaks tonni.
- ♦ 221 triljonit transaktsioonilist kirjet.
- ♦ 100 miljardit struktureerimata dokumenti (sh e-mailid, SMSid ja pildid).
- ♦ Info 30 miljardist (!) allikast nagu kasutajad, sensorid, mobiilsed seadmed.

Kokkuvõte

- ♦ Andmemahud suured
- ♦ Andmemahtude kasv kiire (kuni 50% aastas)
 - Seaduste/määrustega sätestatud kohustus säilitada andmeid mingi ajaperioodi jooksul (näiteks 7 aastat).
 - Trendide, peidetud seoste leidmiseks on vaja andmeid võimalikult pika perioodi kohta.

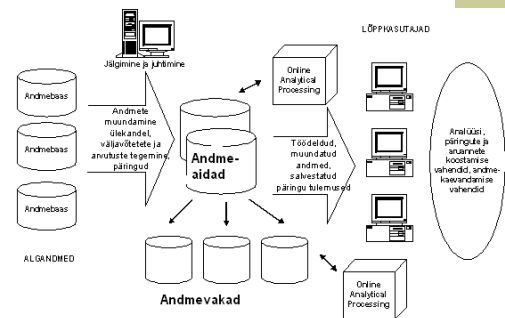
Kokkuvõte (2)

- ♦ Üha rohkem tekib andmeid, mida on potentsiaalselt kasulik säilitada ja analüüsida. Üha rohkem:
 - andmeid sotsiaalvõrgustikest,
 - andmeid veebiliikluse kohta,
 - andmeid sensoritelt e anduritelt.
 - Seade, mis mõõdab või detekteerib reaalses maailmas asetleidvaid protsesse, näit. liikumist, soojust või valgust ja muundab need analoog- või digitaalsignaaleideks

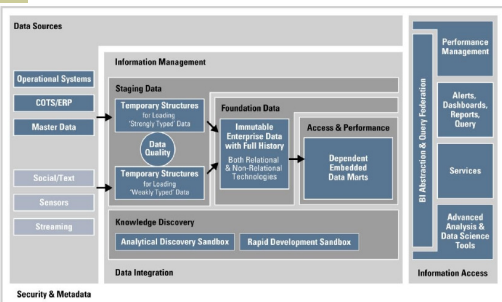
Kokkuvõte (3)

- ♦ Andmete Internetis kättesaadavus paraneb. Andmete kasutamine genereerib omakorda andmeid, mida on kasulik säilitada ja analüüsida.

Andmeida infosüsteemi funktsionaalse arhitektuuri näide



Oracle infohalduse referentsarhitektuur



Allikas: Information Management and Big Data. A Reference Architecture. An Oracle White Paper, February 2013.

Protsessid andmeida infosüsteemis – ETL

- ♦ Andmete **väljalugemine** (eraldamine, hõivamine) operatiivandmete andmebaasidest ja muudest andmeallikatest.
- ♦ Andmeallikatest loetud andmete **kvaliteedi parandamine**
 - puuduvate andmete korral andmeidas kasutatavate andmeväärtuste leidmine,
 - kirjavigade parandamine,
 - standardsete lühendite ja formaatide kasutuselevõtmine,
 - vigaste andmete otsimine, parandamine, väljaajamine,
 - ...
 - **Halva kvaliteediga** andmed lükatakse tagasi – neid ei laadita andmeida.

Protsessid andmeaida infosüsteemis – ETL (2)

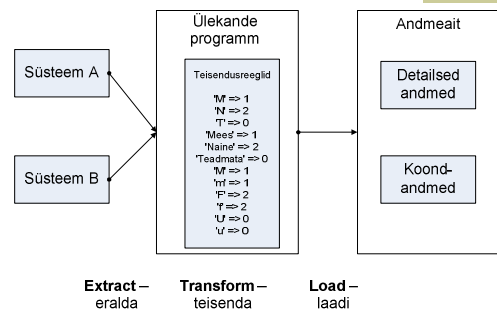
- Andmete **konsolideerimine**.
 - Erinevatest allikatest tulnud andmete põhjal ühtlustatud "pildi" loomine.
 - Näide: Sama mõiste võib erinevates andmebaasides olla tähistatud erinevate andmeväärtustega. Näiteks "N" ja "M"; 1 ja 2; "M" ja "F".
- Sobiliku struktuuriga **failide** loomine.
- Andmete andmeaita **laadimine**.
 - Andmete liigutamine.
 - Kitsenduste kontrollimine.
 - Näide: Unikaalsuse kitsenduse täidetust saab kontrollida andmete laadimise käigus või järel.
 - Indeksi värskendamine.

25.12.2017

Teema 14

43

ETL – kokkuvõte



25.12.2017

Teema 14

44

Ettevalmistusala

- Andmete laadimiseelne töötlemine võib toimuda spetsiaalses **ettevalmistusalas** (eraldi andmebaas).
- Sinna koondatakse andmed erinevatest andmeallikatest.
 - Andmed erinevatest allikatest võivad saabuda erineval ajal.
- Töödeldud andmed laaditakse korraga andmeaita ja eemaldatakse ettevalmistusala.

25.12.2017

Teema 14

45

Big data – suurandmed

- Allikad: Veebiliiklus (logid), sotsiaalvõrgustikud, sensorid (asjade internet), teaduslikud mõõtmised (genoomi järjestamine, astronoomia, ...) ...
- Andmehulgad, mille töötlemine käib olemasolevatele andmebaasisüsteemidele üle jõu.
- Andmeid on *liiga palju*, neid tekib *liiga kiiresti* või nende *struktuur ei sobitu* olemasolevate andmebaaside struktuuriga.
- Töötlemiseks tuleb otsida alternatiivseid võimalusi.

25.12.2017

Teema 14

46

Näiteid suurandmetest

- Raadioteleskoopide massiiv *Square Kilometre Array* hakkab aastas tootma umbes 1 eksabaiti (EB) andmeid (üks miljon terabaiti)
- Kahe miljardi inimese genoomi järjestus – kuni 40 EB
- Youtube – prognoos et aastaks 2025 tekib aastas juurde 1–2 EB andmeid

Allikas: <http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195>

25.12.2017

Teema 14

47

Ettevalmistusala (2)

- Ülisuuri (halvasti struktureeritud) andmehulki (*big data*) võib andmeaita laadimiseks ette valmistada süsteemis, kus on kasutusel *Apache Hadoop* tarkvara raamistik.
- Hadoop* võimaldab andmete hajutatud salvestamist riistvara klastril ning hajutatud töötlemist kasutades *MapReduce* algoritmi.
- Hästi skaleeritav andmetöötlemise süsteem.

25.12.2017

Teema 14

48

Andmebaasid II 2017 © Erki Eessaar

MapReduce kontseptuaalne kirjeldus

25.12.2017 Teema 14 49

Andmebaasid II 2017 © Erki Eessaar

Erinevat tüüpi andmeid – andmete laadimise sagedus

- Andmeita tuleb andmeid juurde laadida, sest allikateks olevad süsteemid "toodavad" pidevalt uusi andmeid.
- "Traditsiooniline" andmeid – andmeid laaditakse (ETL protsesse viiakse läbi) **perioodiliselt**.

25.12.2017 Teema 14 50

Andmebaasid II 2017 © Erki Eessaar

Erinevat tüüpi andmeid – andmete laadimise sagedus (2)

- Reaalaja andmeid** – andmeid laaditakse andmeita **pidevalt**, kohe kui need tekivad.
- Kombinatsioon "traditsioonilisest" ja "reaalaja" andmeidast – osade andmete puhul toimub laadimine pidevalt, osade andmete puhul aga toimub laadimine perioodiliselt.

25.12.2017 Teema 14 51

Andmebaasid II 2017 © Erki Eessaar

Protsessid andmeida infosüsteemis (3)

- Andmete lõppkasutajatele **kättesaadavaks** muutmine.
 - Kasutajal peab olema võimalik kasutada päringute koostamiseks tarkvara, mis tegeleb andmete põhjal koondtulemuste väljaarvutamise ja nende viimise kasutajale sobivasse formaati (risttabelid, graafikud, animatsioonid jne)
 - Kasutajal võib ka olla võimalus registreerida teda huvitavad andmed, et andmeida IS saaks kasutajat teavitada nende andmete muutumisest.

25.12.2017 Teema 14 52

Andmebaasid II 2017 © Erki Eessaar

Protsessid andmeida infosüsteemis (4)

- Andmete **arhiveerimine**.
 - Andmete kirjutamine odavamale, pikemaajaliseks säilitamiseks mõeldud andmekandjale, mis muudab need andmed ka raskemini kättesaadavaks.

25.12.2017 Teema 14 53

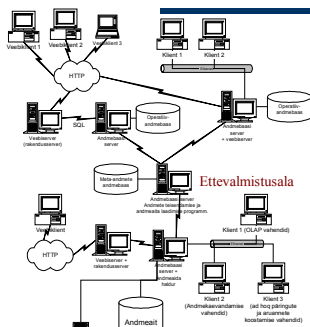
Andmebaasid II 2017 © Erki Eessaar

Andmete lõppkasutajale kättesaadavaks muutmine

25.12.2017 Teema 14 54

Andmebaasid II 2017 © Erki Eessaar

Andmeida infosüsteemi tehnilise arhitektuuri näide



Andmeaitade ja andmevakkade realiseerimiseks kasutatakse enamasti SQL-andmebaasisüsteeme.

Mõnda andmebaasisüsteemi (nt Oracle Enterprise Edition) on andmeida süsteemi loomise vahendid sisse ehitatud.

Ettevalmistusala

25.12.2017 Teema 14 55

Andmebaasid II 2017 © Erki Eessaar

Metaandmed

- Andmeaitade ja andmevakkade süsteem on **metaandmetega** juhitav.
- Metaandmed on andmed andmete kohta.
- Näited:
 - millisest allikast andmed võtta,
 - milliseid teisendusoperatsioone teha,
 - millise andmeida tabeli millisesse veergu andmed salvestada,
 - millised andmed pakuvad huvi millistele kasutajatele.

25.12.2017 Teema 14 56

Andmebaasid II 2017 © Erki Eessaar

Andmeaitade loomise probleeme

- Andmeida andmebaasi skeem (globaalne skeem) peab arvestama kõikide erinevate (nt sadade!) allikatega, kust andmed tulevad.
 - Loomine ja värske hoidmine töö ja ajamahukas
- Andmete kvaliteet
 - Andmeaita tulevad andmed võivad olla valed
 - Võib olla raske saada aru, millised andmed käivad samade ja millised erinevate objektide kohta

25.12.2017 Teema 14 57

Andmebaasid II 2017 © Erki Eessaar

Andmeaitade loomise probleeme (2)



- Allikas 1:* Restoran Vana Karu, Tallinn Kuninga tee 12-3
- Allikas 2:* Restoran Vana Kärü, Tallinn Kuninga tee 12-3
 - Võib-olla üks ja sama restoran
 - Võib-olla kaks restorani, mis tegutsesid samal ajal samas kohas
 - Võib-olla kaks restorani, mis tegutsesid erineval ajal samas kohas

25.12.2017 Teema 14 58

Andmebaasid II 2017 © Erki Eessaar

Ait vs. vakk



Ait

Vakk

25.12.2017 Teema 14 59

Andmebaasid II 2017 © Erki Eessaar

Andmevakk

- Ait** on koht, kuhu asjad säilitamiseks **ära pannakse**.
- Vakk** on mugav koht asjade **otsimiseks** ja **leidmiseks**.
- Luuakse lähtuvalt organisatsiooni üksiku **allüksuse** nõudmistest ja sisaldab andmeid mingi ühe kindla **valdkonna** kohta.
- Andmevakas on üldjuhul **vähem andmeid**, kui andmeidas.

25.12.2017 Teema 14 60

Andmevakk (2)

- Andmed sellesse laaditakse sama allüksuse operatiivandmete andmebaasidest või üleorganisatsioonilisest andmeidast.
- Lõppkasutajad võivad andmevakas olevaid andmeid **uuendada** – näiteks andmekaevandamise käigus ennustuste koostamisel.

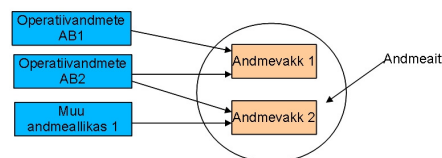
25.12.2017

Teema 14

61

Andmeaida loomise erinevad käsitlused

- Ralph Kimball** (*alt-üles lähenemine*): Kõigepealt luuakse andmevakad. Andmevakas olevad tabelid on tugevasti denormaliseeritud. Andmeait on andmevakkade kogum.



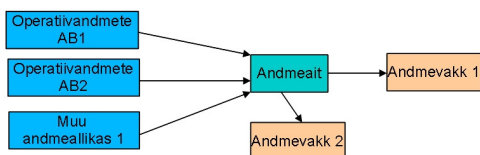
25.12.2017

Teema 14

62

Andmeaida loomise erinevad käsitlused (2)

- Bill Inmon** (*ülalt-alla lähenemine*): Kõigepealt luuakse andmeait ja seejärel andmevakad, mis saavad oma andmeid andmeidast. Andmeidas on tabelid vähemalt kolmandal normaalkujul.



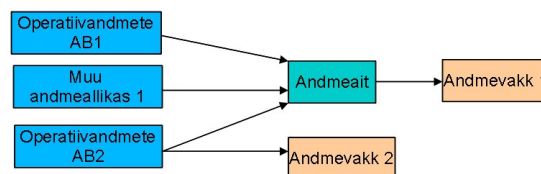
25.12.2017

Teema 14

63

Andmeaida loomise erinevad käsitlused (3)

- Kombineeritud lahendus.



25.12.2017

Teema 14

64

Andmeaida andmebaasi disain

- Järgnev andmeaida disaini kirjeldus põhineb *Bill Inmoni* käsitlusel.
- Aluseks võetakse operatiivandmete andmebaaside disaini kirjeldus.

25.12.2017

Teema 14

65

Andmeaida andmebaasi disain (2)

- Andmeaida **mustrite** otsimine.
- Eemalda andmed, mida andmeaida põhjal tehtavates päringutes ei kasutata.
- Tabelite **primaarvõtmetesse** uue(te) veer(gu)(de) lisamine (ajaandmed).

| isik_id | eesnimi | perenimi | algus | lopp |
|---------|--------------|----------|------------|------------|
| 2 | Antuan | Thury | 01.01.1981 | 05.02.1999 |
| 2 | Anton | Turi | 06.02.1999 | 17.02.2000 |
| 2 | Anton-Andres | Turi | 18.02.2000 | 31.12.9999 |

Primaarvõti: (isik_id, algus, lopp)

25.12.2017

Teema 14

66

Andmeaida andmebaasi disain (3)

- ♦ Andmete **granulaarsuse** (teralisuse, detailsuse) valimine
 - Näide: Kas rida tabelis *Arve* vastab ühele konkreetsele arvele, ühe päeva kõigile arvetele, ühe nädala kõigile arvetele?
- ♦ **Seosetüüpidele** vastavate tabelite lisamine
 - Näide: Igal ajahetkel projektil üks projektijuht, kuid ajaloolises plaanis võib projektil olla mitu projektijuhti – erinevatel aegadel.

25.12.2017

Teema 14

67

Andmeaida andmebaasi disain (4)

- ♦ **Töökiiruse** huvidest tulenevad disainimuudatused
 - Tuletatud andmete sissetoomine
 - Näide: Arve summa, mille saab arvutada arve ridade põhjal
 - Tabelite valikuline denormaliseerimine
 - Tabelite valikuline ühendamine
 - Veergude gruppide lisamine tabelitesse
 - Näide: Myyk (myyja, jaanuar, veebruar, marts, aprill, mai,)

25.12.2017

Teema 14

68

Andmeaida andmebaasi disain (5)

- ♦ Tabelite teisendamine nii, et ühes tabelis olevatele veergudel on suhteliselt ühesugune muutumise kiirus (sisuliselt *vertikaalne killustamine*):
- ♦ *Klient_muutumatu*(klient_id, kuupäev, synni_aeg) Primaarvõti (klient_id, kuupäev)
- ♦ *Klient_muutub_harva*(klient_id, kuupäev, eesnimi, perenimi) Primaarvõti (klient_id, kuupäev)
- ♦ *Klient_muutub_sageli*(klient_id, kuupäev, krediitireiting, seisund) Primaarvõti (klient_id, kuupäev)

25.12.2017

Teema 14

69

Andmeaida andmebaasi disain (6)

- ♦ Tabelitele **kitsenduste** lisamine
 - Võimaldavad andmebaasisüsteemil andmetest "aru saada"
 - Aitavad andmebaasisüsteemi optimeerimismoodulil valida parima täitmisplaani (semantiline teisendamine)

25.12.2017

Teema 14

70

Andmeaida andmebaasi disain – metaandmed

- ♦ Sellised andmed võimaldavad jälgida andmete laadimise protsessi ja andmete kvaliteeti ning vajadusel andmete laadimine tagasirullida (tühistada).

25.12.2017

Teema 14

71

Andmeaida andmebaasi disain – metaandmed (2)

- ♦ Kaaluda tabelitesse täiendavate veergude lisamist.
 - Rea andmete operatiivandmete andmebaasist eraldamise aeg
 - Rea andmete laadimise aeg
 - Laadimistsükkel
 - Kas tegu on hetkel kehtivate andmetega?
 - Süsteem, kust kohast on andmed pärit
 - Andmete usaldusväärsuse hinnang

25.12.2017

Teema 14

72

Andmebaasid II 2017 © Erki Eessaar

Metaandmed – laadimistsükkel

| laadimistsykli_id | laadimise_aeg | laadija_id | seisund |
|-------------------|---------------|------------|-----------|
| 5 | 01-Jun-2002 | 34 | lõpetatud |
| 4 | 01-May-2002 | 34 | lõpetatud |
| 3 | 01-Apr-2002 | 35 | lõpetatud |
| 2 | 01-Mar-2002 | 34 | lõpetatud |
| 1 | 01-Feb-2002 | 34 | lõpetatud |

25.12.2017 Teema 14 73

Andmebaasid II 2017 © Erki Eessaar

Metaandmed – andmete päritolu

| allika_id | nimetus |
|-----------|---------------------------|
| 1 | Tellimuste andmebaas |
| 2 | Raamatupidamise andmebaas |
| 3 | Klientide andmebaas |
| 4 | Andmeait X |
| 5 | Internet |

25.12.2017 Teema 14 74

Andmebaasid II 2017 © Erki Eessaar

Metaandmed – usaldusväarsuse hinnang

| usaldustaseme_id | kirjeldus |
|------------------|--|
| 5 | Otse operatiivandmete andmebaasist. |
| 4 | Saadud mitmest operatiivandmete andmebaasist pärit andmete integreerimise tulemusel. |
| 3 | Tuletatud |
| 2 | Hinnangul põhinev |
| 1 | Teadmata |

25.12.2017 Teema 14 75

Andmebaasid II 2017 © Erki Eessaar

Veel tegevusi andmeaida disaini käigus

- ♦ Võimalike **andmeallikate** leidmine
- ♦ Andmete **profileerimine** – andmete **kvaliteedi** hindamine allikates
 - Enne disaini algust, otsustamaks kas on üldse mõistlik andmeaida loomisega hetkel alustada
 - Disaini käigus
 - sobivate allikate valimiseks
 - otsustamaks, kas vajavad eelnevat kvaliteedi tõstmist
 - andmaks sisendit tõlkimis ja integreerimisreeglitele

25.12.2017 Teema 14 76

Andmebaasid II 2017 © Erki Eessaar

Veel tegevusi andmeaida disaini käigus (2)

- ♦ **Tõlkimis-** ja **integreerimisreeglistiku** loomine
- ♦ Andmete **laadimise sageduse** valimine
- ♦ **Turvasüsteemi** projekteerimine
- ♦ Andmeaida loomise **vahendite** valimine
- ♦ Andmeaida **tehnilise arhitektuuri** kavandamine
- ♦ Andmeaida andmebaasi **füüsiline disain**

25.12.2017 Teema 14 77

Andmebaasid II 2017 © Erki Eessaar

Andmeaida eeliseid

- ♦ Paremad andmed otsuste langetamiseks.
 - Andmed kokku kogutud, integreeritud, nende põhjal on lihtsam päringuid teha.
- ♦ **Onlain-tehingutöötluste** süsteemid ja **analüüsi/otsustussüsteemid** ei koorma oma tööga üksteist.
- ♦ Võimalus integreerida tulevikus uusi andmeallikaid.
- ♦ Tekib ülevaade andmete kvaliteedist.
 - Samas halb andmete kvaliteet operatiivandmete andmebaasides pikendab süsteemi loomiseks kuluvat aega.

25.12.2017 Teema 14 78

Andmeaida väljakutsed

- ♦ Arvutiressursi vajadus kasvab.
- ♦ Arendus on pikaajaline ja keerukas.
 - Näiteks kuidas laadida kiiresti suuri andmehulki või tagada aktsepteeritav päringute töökiirus.
- ♦ Andmete konfidentsiaalsus võib saada rikutud.
- ♦ Võib tekkida vajadus täiendada operatiivandmete andmebaase.

25.12.2017

Teema 14

79

Andmeaida ja andmevaka ühisomadused

- ♦ Valdavad operatsioonid on:
 - keerukad päringud ja
 - andmete lisamine (laadimise käigus).
- ♦ Kustutamine toimub vaid siis, kui andmemahud muutuvad ülisuureks või seda nõuavad seadusaktid.

25.12.2017

Teema 14

80

Andmeaitade ja andmevakkade kasutamine

- ♦ Andmeaitu ja andmevakke kasutatakse küsimustele vastuste saamiseks, et langetada juhtimisotsuseid.
- ♦ Küsimused võivad olla.
 - *Valmis küsimused*, millele soovitakse vastust iga päev või kuu.
 - *Ad hoc küsimused*, mille peale just tuldi.
 - *Küsimused, mida veel ei osata küsida*. Seda võib nimetada avastamisprotsessiks ehk "ma tean küsimust, kui näen vastust".

25.12.2017

Teema 14

81

Skeemi disaini erinevusi

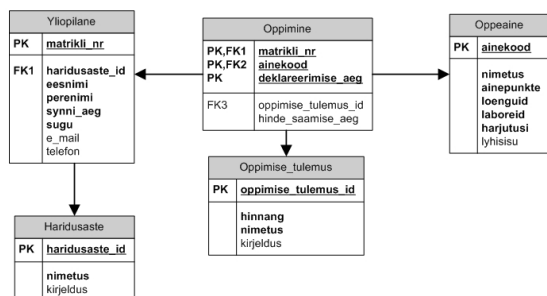
- ♦ Operatiivandmete andmebaasides on tabelid kõrge astmeni normaliseeritud.
- ♦ Andmeaida tabelite struktuur baseerub operatiivandmete andmebaaside struktuuril.
 - Ei koguta andmeid, mille ajalugu ei soovita säilitada.
 - Andmebaasis võivad lisaks olla ka koondandmed, et kiirendada päringutele vastamist (nt tellimuse juures välja arvatud kogumaksumus).
 - Päringukiiruse huvides võivad mõned tabelid olla denormaliseeritud.
- ♦ Andmevakas on tabelid **denormaliseeritud** ning on organiseeritud **täht-** või **lumehelbe** struktuuri järgi.
 - Keskne faktitabel on seotud mitme dimensioonide tabeliga.

25.12.2017

Teema 14

82

Operatiivandmete andmebaasi tabelite disaini näide

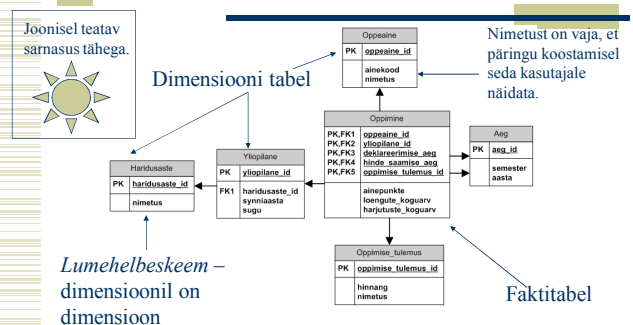


25.12.2017

Teema 14

83

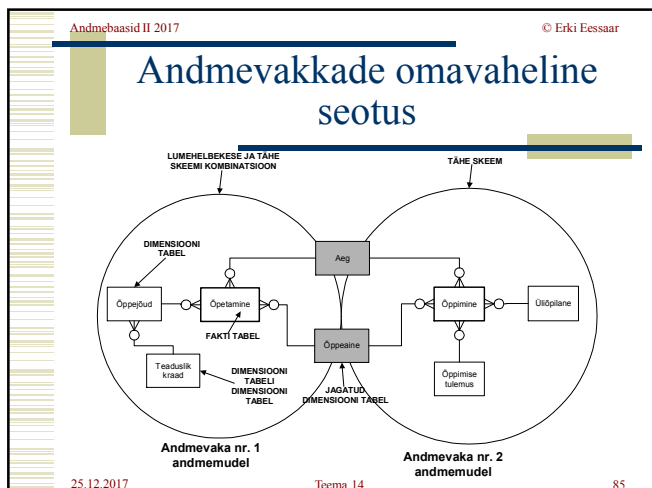
Andmevaka tabelite disaini näide



25.12.2017

Teema 14

84



Andmebaasid II 2017 © Erki Eessaar

Päringud andmevakast

- ♦ *Star query* – faktitabel ühendatakse ühe või mitme dimensioonitabeliga kasutades primaarvõtme/välisvõtme veerge.
 - Mitu naissoost üliõpilast on 2017. aastal õppimist alustanud?
 - Kasutatud dimensioonid: *Yliõpilane, Aeg*
 - Millised on 2017. aastal lõppenud õppimise tulemused? Leidke iga hinde (tulemuse kohta) selle saanud üliõpilaste arv.
 - Millised on 2017. aasta põhjal 10 kõige kõrgema keskmise õppimise tulemusega õppeainet?

25.12.2017 Teema 14 87

Andmebaasid II 2017 © Erki Eessaar

Andmevaka andmebaasi disain

- ♦ Andmevaka valdkonna valimine
- ♦ Andmevaka kirjeldavate mustrite otsimine
- ♦ Faktitabelis hoitavate andmete täpsuse valimine
- ♦ Dimensioonide valimine
- ♦ Faktitabelis hoitavate andmete (veergude) valimine
- ♦ Arvutatavate väärtustele vastavate veergude lisamine faktitabelisse

25.12.2017 Teema 14 88

Andmebaasid II 2017 © Erki Eessaar

Andmevaka andmebaasi disain (2)

- ♦ Dimensiooni tabelitesse täiendavate kommentaari veergude lisamine
- ♦ Andmebaasis hoitavate andmete vanuse valimine
- ♦ Aeglaselt muutuvate dimensioonide leidmine

25.12.2017 Teema 14 89

Andmebaasid II 2017 © Erki Eessaar

Dimensiooni väärtuste muutumise probleemi lahendamine

| õppimise_tulemus_id | hinnang | nimetus | kehtivuse algus |
|---------------------|---------|-------------|-----------------|
| 1 | 0 | puudulik | 01.09.1995 |
| 2 | 1 | nõrk | 01.09.1995 |
| 3 | 2 | kasin | 01.09.1995 |
| 4 | 3 | rahuldav | 01.09.1995 |
| 5 | 4 | hea | 01.09.1995 |
| 6 | 5 | väga hea | 01.09.1995 |
| 7 | 0 | puudulik | 01.09.2000 |
| 8 | 1 | kasin | 01.09.2000 |
| 9 | 2 | rahuldav | 01.09.2000 |
| 10 | 3 | hea | 01.09.2000 |
| 11 | 4 | väga hea | 01.09.2000 |
| 12 | 5 | suurepärase | 01.09.2000 |

Parem

| õppimise_tulemus_id | hinnang | nimetus_1995 | nimetus_2000 |
|---------------------|---------|--------------|--------------|
| 1 | 0 | puudulik | puudulik |
| 2 | 1 | nõrk | kasin |
| 3 | 2 | kasin | rahuldav |
| 4 | 3 | rahuldav | hea |
| 5 | 4 | hea | väga hea |
| 6 | 5 | väga hea | suurepärase |

Halvem – nõuab tabelite struktuuri muutmist.

25.12.2017 Teema 14 90

Vaadake ka raamatuid TTÜ raamatukogust

- ♦ Silverston, L., 2001. *The Data Model Resource Book. A Library of Universal Data Models for All Enterprises*. Revised Edition. Vol. 1. Wiley Computer Publishing. 540 p.
- ♦ Silverston, L., 2001. *The Data Model Resource Book. A Library of Universal Data Models by Industry Types*. Revised Edition. Vol. 2. Wiley Computer Publishing. 556 p.

Tarkavara

- ♦ Kõige olulisemad andmeaitade tarkvara loojad 2017. aasta alguses.

- Oracle
- Teradata
- Microsoft
- IBM
- SAP
- Amazon Web Services

Magic Quadrant for Data Management Solutions for Analytics. February, 2017.

Tarkvara arengusuundi

- ♦ Veerupõhine andmete salvestamine.
 - Kiiremad koondandmete päringud.
 - Väiksem andmemaht tänu paremale pakkimisele.
- ♦ Integratsioon *Apache Hadoop* tarkvara raamistikuga.
 - Andmeid saab küsida SQLi abil.
- ♦ Andmeait kui pilvandmetöötluse teenus.
- ♦ Kogu andmebaasi hoidmine muutmälus.

Loogiline andmeait

- ♦ Andmeid ei koondata ühte võrgu sõlme kokku, vaid need jäävad andmebaasi kuhu need algselt lisati
 - Andmebaasid võivad olla heterogeensed, sealhulgas *Hadoopil* põhinevad
 - Andmeid, mida liigutada tekib liiga palju ja liiga kiiresti – parem jätta need sinna kus need tekkisid

Loogiline andmeait (2)

- ♦ Tarkvara võimaldab näha neid andmeid loogilise tervikuna
- ♦ Sarnane teemas 12 kirjeldatud *föderatiivsele süsteemile*
- ♦ Võimalik, et osa andmeid kantakse "traditsioonilisse" andmeaita/ andmevakkadesse, mis on omakorda loogilise andmeaita osad
- ♦ Odavam, kiirem arendada
 - Üks võimalus kasutada PostgreSQLi väliseid tabeleid




Andmejärv

- ♦ Millal loomist kaaluda?
 - Suurandmete korral – kui andmeid hästi palju ja hästi erineva struktuuriga
- ♦ Erinevusi andmeaitadest?
 - Andmeid ei töödelda enne laadimist ja ei panda ettevalmistatud andmestruktuuridesse
 - Selle asemel salvestatakse andmed selles formaadis nagu need tekkisid, lisades juurde metaandmed (nt info tekkekoha kohta)

Andmebaasid II 2017 © Erki Eessaar

Andmejärv (2)



- Andmejärv "kalastavad" *andmeteadlased*, kes otsivad sealt oma meetoditega peidetud seoseid trende jne
 - Andmeaitade ja andmevakkadele võivad erinevate analüüsi ja päringusüsteemidega ligi pääseda ka *ärikasutajad*
 - Andmejärv ei oskaks nad midagi peale hakata
- Andmejärv võib kergesti muutuda *andmesooks*, kust väga raske midagi leida

25.12.2017 Teema 14 97

Andmebaasid II 2017 © Erki Eessaar

Andmesoo analoog pärismaailmas



25.12.2017 Teema 14 98

Andmebaasid II 2017 © Erki Eessaar

Andmejärv (3)

- Füüsiline andmejärv* – andmed kopeeritakse realselt ühte kohta (serverite kobarasse) kokku
- Loogiline andmejärv* – andmed jäävad sinna kus tekkisid, kuid pealisehitiseks olev virtuaalne andmete kiht pakub neist tervikpildi

25.12.2017 Teema 14 99

Andmebaasid II 2017 © Erki Eessaar

Oracle vahendid andmeaida/vaka loomiseks

- SQL Loader
 - Vahend andmete kiireks laadimiseks andmebaasi välistest failidest Oracle andmebaasi.
- Oracle Warehouse Builder (OWB)
 - Pakub graafilise kasutajaliidese Oracle andmebaasisüsteemil põhinevate andmeaitade ja andmevakkade loomiseks. Võimaldab näiteks:
 - disainida ETL protsesse,
 - modelleerida andmeaida/andmevaka andmebaasi.

25.12.2017 Teema 14 100

Andmebaasid II 2017 © Erki Eessaar

Oracle vahendid andmeaida/vaka loomiseks (2)

- Hetktõmmised – kiiremad vastused päringutele
- Virtuaalsed veerud – abistavad koondandmete esitamisel
- Bitmap ja bitmap-join indeksid
- Päringu sisemise taseme täitmismeetod – *star transformation*
- Välised tabelid – andmete laadimine
- Tabelite sektsioonideks jagamine
- Andmete pakkimine

25.12.2017 Teema 14 101

Andmebaasid II 2017 © Erki Eessaar

Oracle – lisavõimalused andmekäitluskeeles lausete kirjutamiseks


- MERGE lause (lihtsustab andmete laadimist)
- Analüütilised funktsioonid (aknafunktsioonid)
- Grupeerimine – ROLLUP, CUBE, GROUPING SETS
- MODEL klausel
- Päringud hierarhiliste andmete põhjal
- PIVOT ja UNPIVOT operaatorid
- Paralleeltöö SQL lausete täitmisel
- Konveiertööga tabelifunktsioonid

25.12.2017 Teema 14 102

Andmebaasid II 2017 © Erki Eessaar

Andmeaitu ja andmevakke kasutavad programmid

- ♦ Pääringusüsteemid
 - Tavakeeles tehtavad pääringud
- ♦ OLAP – Online Analytical Processing – reaajas analüütiline töötlemise vahendid
- ♦ Otsustussüsteemid
- ♦ Andmekavandamise programmid
 - "Peidetud" seoste otsimine, prognooside tegemine



25.12.2017 Teema 14 103

Andmebaasid II 2017 © Erki Eessaar

Toorandmed vs. kokkuvõtted




- ♦ Inimesel on suurt hulka toorandmeid raske hoomata.
- ♦ Ta soovib nende põhjalt tehtud kokkuvõtteid.
 - **Vahemärkus!** Pole õige, et kasutusjuht "X koondaruande vaatamine" tähendab lihtsalt kõikide X eksemplaride andmete vaatamist.

25.12.2017 Teema 14 104

Andmebaasid II 2017 © Erki Eessaar

OLTP vs. OLAP

- ♦ OLTP süsteem – Onlain tehingutötluse süsteem
- ♦ OLAP süsteem – Onlain analüüsi süsteem

25.12.2017 Teema 14 105

Andmebaasid II 2017 © Erki Eessaar

OLTP vs. OLAP süsteemi kasutajad

| OLTP süsteemi kasutaja (ametnik, teenindaja) | OLAP süsteemi kasutaja (analüütik, juhataja) |
|--|---|
| Andmete sisestamine, kontrollimine | Andmete uurimine, prognooside tegemine |
| Vajab enamasti hetkel kehtivaid andmeid | Vajab ajaloolisi ja hetkel kehtivaid andmeid |
| Vajab vastust kohe | Vajab vastust kohe, kuid äärmisel juhul võib veidi oodata |

25.12.2017 Teema 14 106

Andmebaasid II 2017 © Erki Eessaar

Nõudmised OLAP süsteemile – FASMI

- ♦ Fast – kiire
- ♦ Analysis – analüüs
- ♦ Shared – jagatud
 - Erinevad kasutajad peaks saama pöörduda samade andmete poole
- ♦ Multidimensional – mitmemõõtmeline
- ♦ Information – informatsioon
 - Analüüsi tulemuseks on kasulik informatsioon

25.12.2017 Teema 14 107

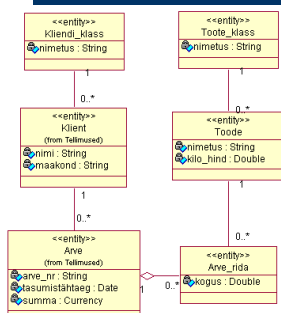
Andmebaasid II 2017 © Erki Eessaar

OLAP vahendite tüübid

- ♦ MOLAP – Mitmemõõtmeline OLAP.
 - Andmed on mitmemõõtmelises andmebaasis.
 - Selles on andmed salvestatud mitmemõõtmelise massiivi pesadesse.
 - Pääringute vastused on ette valmis arvutatud.
- ♦ ROLAP – Relatsiooniline OLAP.
 - Andmed on relatsioonilises (SQL) andmebaasis.
- ♦ HOLAP – Hübriidne OLAP.
 - Koondandmed mitmemõõtmelises andmebaasis, detailsed andmed relatsioonilises andmebaasis

25.12.2017 Teema 14 108

Operatiivandmete andmebaas



25.12.2017

Teema 14

109

Sõltuvad- ja sõltumatud muutujad mitmemõõtmelises andmebaasis

- ♦ **Dimensioonid** (sõltumatud muutujad) on andmete erinevad aspektid, mida kasutatakse andmete jagamiseks kindlat tähendust omavatesse rühmadesse.
- ♦ **Sõltuvad muutujad** on *arvulised* mõõdud.

25.12.2017

Teema 14

110

Päringud mitmemõõtmelisest andmebaasist

- ♦ Näidake viimase **viie aasta** jooksul teenitud **tulu** erinevate **tooteklasside** lõikes.
 - **Aeg**, **tooteklass** – dimensioonid
 - **Tulu** (**rahasumma**) – mõõt
- ♦ Näidake viimase kümne aasta jooksul läbi viidud müükide arvu erinevate kliendiklasside lõikes.
- ♦ Näidake erinevate kliendiklasside lõikes, kui palju nad on viimase viie aasta jooksul mingisse tooteklassi kuulunud tooteid ostnud.

25.12.2017

Teema 14

111

Mitmemõõtmeline andmebaas – näide kahe dimensiooni lõikest

Dimensioonide
väärtused

Mitmemõõtmeline massiiv

| | Hulgi | Jae | Eksport | Kokku |
|-----------|-------|-----|---------|-------|
| Halvaa | 50 | 60 | 100 | 210 |
| Shokolaad | 40 | 70 | 80 | 190 |
| Kommid | 90 | 120 | 140 | 350 |
| Näts | 20 | 10 | 30 | 60 |
| Kokku | 200 | 270 | 350 | 810 |

Ette väljaarvutatud koondtulemused
(päringute vastused)

Mõõdu
väärtused

25.12.2017

Teema 14

112

Päringu tegemine

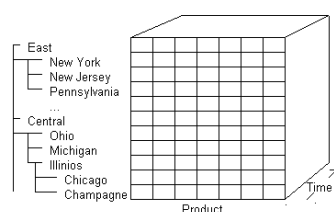
- ♦ Vali dimensioonid
 - Näide: aeg, tooteklass
- ♦ Vali dimensioonile rakendatavad piirangud
 - Näide: aeg: aastad 2000–2004
- ♦ Vali mõõdu väärtustele rakendatav kokkuvõttefunktsioon (Sum, Count, Avg, ...)
 - Näide: leia iga tooteklassi ja aasta kohta müüdud kauba koguse summa (funktsioon *Sum*)

25.12.2017

Teema 14

113

Hierarhilised dimensioonid



Veel hierarhiate
näiteid:

Riik

Maakond

Vald

Küla

Linn

Tänav

25.12.2017

Teema 14

114

Andmebaasid II 2017 © Erki Eessaar

Aeg kui hierarhiline dimensioon

...

Aasta
 Poolaasta
 Kvartal
 Kuu
 Päev
 Tund
 Minut
 Sekund

25.12.2017 Teema 14 115

Andmebaasid II 2017 © Erki Eessaar

Hierarhilised dimensioonid (2)

| | 2000 | | 2001 | |
|-----------------|-------|-------|-------|-------|
| | North | South | North | South |
| VCR Type1 | 100 | 150 | 150 | 100 |
| VCR Type2 | 200 | 250 | 250 | 200 |
| VCR Type3 | 150 | 100 | 150 | 150 |
| Camcorder Type1 | 400 | 400 | 400 | 450 |
| Camcorder Type2 | 500 | 600 | 550 | 600 |

Drill-up ↓ ↑ Drill-down

| | 2000 | | 2001 | |
|-----------|-------|-------|-------|-------|
| | North | South | North | South |
| VCR | 450 | 500 | 550 | 450 |
| Camcorder | 900 | 1000 | 950 | 1050 |

25.12.2017 Teema 14 116

Andmebaasid II 2017 © Erki Eessaar

Hõre mitmemõõtmeline andmebaas

- 3 dimensiooni – toode, regioon, aeg
- 100 toodet, 100 regiooni, 52 nädalat
- Andmebaasis on $100 \times 100 \times 52 = 520000$ lahtrit.
- Juhul kui **vähem kui 10%** andmebaasi lahtreid sisaldab andmeid öeldakse, et andmebaas on **hõre**
- Olemas spetsiaalsed meetodid andmebaasi **pakkimiseks**

25.12.2017 Teema 14 117

Andmebaasid II 2017 © Erki Eessaar

Hõre mitmemõõtmeline andmebaas (2)

- Andmeväärtuse puudumine mitmemõõtmelise massiivi lahtris võib tähistada erinevaid olukordi.
 - Andmed on teadmata.
 - Mõnda toodet polegi enamikes regioonides müüdud.
 - Mõni toode tuli müügile alles viimasel kahel nädalal.
 - Andmed on olemas, aga neid pole baasi kantud.
 - Andmed on ebasobivad.

25.12.2017 Teema 14 118

Andmebaasid II 2017 © Erki Eessaar

ROLAP

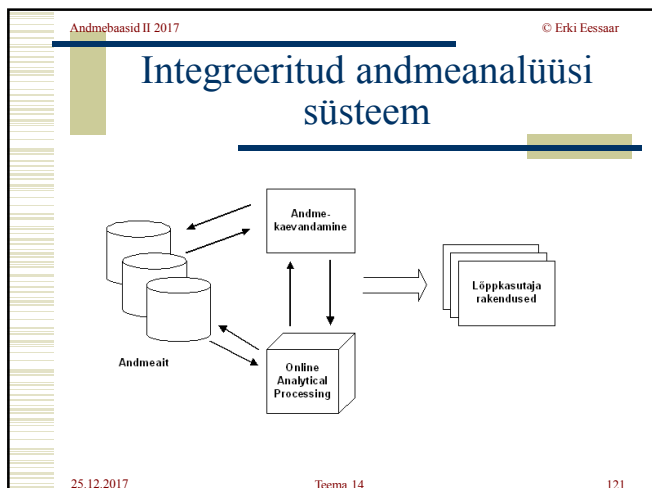
Andmete küsimine relatsioonilisest (SQL) andmebaasist koormab andmebaasisüsteemi. Pärangule vastuse leidmine võtab aega. Kõikvõimalikke pärangu tulemusi pole ette välja arvatud. Sellist süsteemi aitavad realiseerida CUBE ja ROLLUP klauslid SQL lausetes.

25.12.2017 Teema 14 119

Andmebaasid II 2017 © Erki Eessaar

Managed Query Environment (päringukeskkond)

25.12.2017 Teema 14 120



Andmebaasid II 2017 © Erki Eessaar

Kuidas teha *kiiresti* kokkuvõtteid, kui andmeid palju?

NEED FOR SPEED

- ♦ Arvutada *vahetulemused ette valmis*, et nende pealt kiiresti koondtulemus produtseerida.
 - Seda lahendust kasutavad MOLAP ja HOLAP süsteemid.

25.12.2017 Teema 14 122

Andmebaasid II 2017 © Erki Eessaar

Kuidas teha *kiiresti* kokkuvõtteid, kui andmeid palju? (2)

- ♦ Kui andmed SQL-andmebaasis *veerupõhiselt salvestatud*, siis koondandmete arvutamine läheb palju kiiremini (võrreldes reapõhise salvestusega), sest süsteemil on vaja läbi vaadata vähem andmeid.
 - Sellest saaks kasu ROLAP.
 - Kui andmebaasisüsteem pakuks hea kasutajaliidese selliste päringute tulemuste vaatamiseks, poleks ehk eraldi OLAP programmi vajagi.

25.12.2017 Teema 14 123

Andmebaasid II 2017 © Erki Eessaar

Kuidas teha *kiiresti* kokkuvõtteid, kui andmeid palju? (3)

- ♦ Teha kokkuvõte mingi *valimi* põhjal.
 - Läbi ei vaadata kõiki andmeid, vaid ainult osa.
 - Tuleb leida hea valim, et tulemus oleks võimalikult täpne.
 - Ka sellest võiks ROLAP kasu saada.

25.12.2017 Teema 14 124

Andmebaasid II 2017 © Erki Eessaar

Ankurmodelleerimine (Anchor Modeling)

- ♦ Modelleerimistehnika, mis võimaldab **agilset** andmebaasi projekteerimist.
- ♦ Selle tehnika toetuseks on loodud avatud lähtekoodiga veebipõhine modelleerimisvahend ning koodigeneraator.
- ♦ <http://www.anchor modeling.com/>
- ♦ Saab kasutada nii **andmeaitade** kui ka operatiivandmete andmebaaside loomiseks.

25.12.2017 Teema 14 125

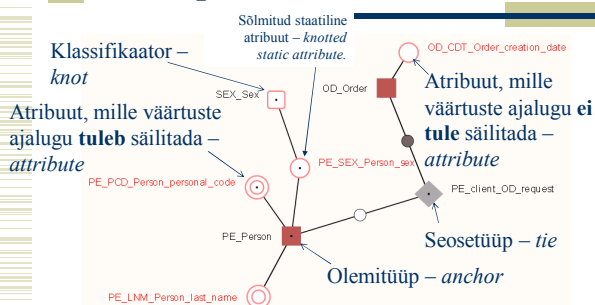
Andmebaasid II 2017 © Erki Eessaar

Ankurmodelleerimine (2)

- ♦ Modelleerimiskeel võimaldab kirjeldada olemitüüpe (*anchors*), atribuute (*attributes*), seosetüüpe (*ties*) ja klassifikaatoreid (*knots*)
 - Keele konkreetne süntaks meenutab Chen'i ERD süntaksi.
- ♦ Saab määrata atribuudid, seosetüübid ja klassifikaatorid, mille korral soovitakse säilitada **andmete ajalugu**.
- ♦ Loodud mudelist saab genereerida **SQL laused** andmebaasi loomiseks.

25.12.2017 Teema 14 126

Ankurmodelleerimine – mudel veebipõhises CASE vahendis



<http://www.anchormodeling.com/>

25.12.2017

Teema 14

127

6 Ankurmodelleerimine (3)

- ◆ Kasutatakse *teisendusreeglid*, mis loovad mudeli elementidest üks-ühele tabelid – tulemuseks on **kuuendal normaalkujul** olevad tabelid.
 - Võimaldab vältida NULLide kasutamist puuduvate andmete esitamiseks.
 - Võimaldab efektiivselt hallata andmete ajalugu.
- ◆ Olemitüüpide, atribuutide, klassifikaatorite põhjal tekitab kuuendal normaalkujul tabelid.
- ◆ Seosetüüpide põhjal tekkivad tabelid võivad olla madalamal normaalkujul.

25.12.2017

Teema 14

128

Ankurmodelleerimine – tabelite loomise laused

```

-- Sex_Sex table
--
IF NOT EXISTS (SELECT * FROM sys.objects WHERE name = 'Sex_Sex' and type LIKE 'IU')
CREATE TABLE [Sex_Sex] (
    SEX_ID SMALLINT not null,
    SEX VARCHAR(10) not null unique,
    Metadata PK int not null,
    primary key (
        SEX_ID and
    )
);
GO

-- EP_Person table
--
IF NOT EXISTS (SELECT * FROM sys.objects WHERE name = 'EP_Person' and type LIKE 'IU')
CREATE TABLE [EP_Person] (
    EP_ID int identity(1, 1) not null,
    Metadata PK int not null,
    primary key (
        EP_ID and
    )
);
GO

```

Sujuv, üks-ühele teisendus tabeliteks

**Enamik tabeleid
kuuendal
normaalkuiul**

Võimalik säilitada
andmete ajalugu

Skeemi evolutsio-
neerimine tähendab uute
tabelite lisamist (vähem
probleeme
andmesiirdega vanast
skeemi versioonist uude)

25.12.2017

Teema 14

129

Ankurmodelleerimine (4)

- ◆ Lisaks tabelite loomise lausetele genereeritakse ka laused loomaks **vaated**, mis esitavad andmeid "traditsioonilise disainiga" tabelitena, kus on ühe olemitüübi erinevatele atribuutidele vastavad andmed.
- ◆ Kui modelleerija avaldab soovi säilitada atribuudi väärtuste ajalugu, siis lisandub atribuudi alusel loodud tabelisse veerg, milles hoitakse andmeid atribuudi väärtuse muutumise aja kohta.

25.12.2017

Teema 14

130

Ankurmodelleerimine – andmebaasi- operatsioonide töökiirus

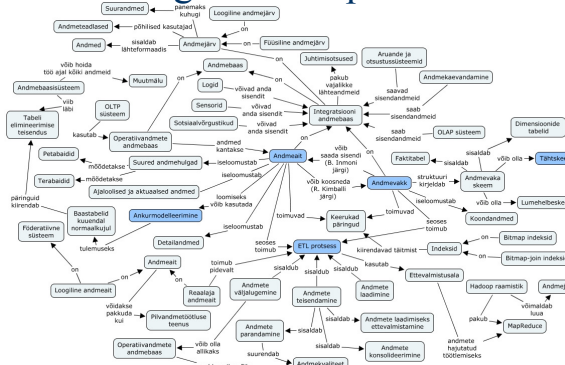
- ◆ Päringute töökiiruse paranemisele aitab kaasa:
 - väike veergude hulk tabelis,
 - mõnede andmebaasisüsteemide (nt PostgreSQL 9.5, Oracle 12c Enterprise Edition Release 1, MS SQL Server, MariaDB) poolt andmekäitluskeele lausete täitmisplaanide optimeerimise käigus rakendatav *tabeli elimineerimise* teisendus (kui päringus viidatud tabeli kasutamine ei mõjuta päringu tulemust, siis võtab andmebaasisüsteem kasutusele sellise täitmisplaan, mille kohaselt selle tabeli andmeid ei loeta),
 - minimaalne andmete liiasus.

25.12.2017

Teema 14

131

Mõningad teema põhimõisted



25.12.2017

Teema 14

132

