Andmeaitade loomise koolitusseminar

08.02.2022

RIIGI INFOSÜSTEEMI AMET

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

TARK

e-RIIK

Päevakava

08.30-09.00 Kogunemine ja tervituskohv

09.00-09.45 Sissejuhatus ja põhimõisted

09.45-10.00 Paus

10.00-11.15 Andmeait/andmeladu

11.15-11.30 Sirutuspaus

11.30-12.00 Andmelao tehnoloogiad ja tarkvarad

12.00-13.00 Lõunapaus

13.00-14.00 Andmelao projekt

14.00-14.15 Paus

14.15-14.45 Andmeladu ja Eestis kehtivad nõuded

14.45-15.00 Küsimused/vastused

Sissejuhatus ja põhimõisted

Riigi Infosüsteemi Amet
TARK e-RIIK

Resta OÜ

- 30 aastat kogemust andmelaonduse ja ärianalüüsi valdkonnas
- Spetsialiseerunud analüüsilahenduste loomisele
  - Andmelaod
  - Aruandlus- ja analüütikalahendused
  - Suurandmete (Big Data) lahendused
  - Andmekvaliteedi analüüsi ja monitoorimise süsteemid
  - Statistilise analüüsi ja andmekaeve keskkondade loomine, koolitused ja konsultatsioonid
  - Finantsanalüüsisüsteemid koos konsolideerimise ja eelarvestamise moodulitega
- 15 spetsialisti

Ain Rääbis

- Tartu Ülikool, Matemaatilise statistika magister
- Restas alates 1997. aastast, ärianalüüsi lahenduste loomine
- Erinevad rollid:
  - Arhitekt
  - Analüütik
  - Arendaja/ programmeerija
  - Projektijuht
  - Andmekvaliteedi spetsialist
  - Koolitaja
- Kogemus paljude andmelao- ja analüüsi-projektidega (TEHIK, PPA/SMIT, Tallink, Omniva, Justiitsministeerium/RIK, Telia andmekaeve, TEHIK/SOM, Töötukassa, EAS, ERGO kindlustus jne)

Vajadus andmelao järele – veidi ajaloost

- 1950ndad aastat – algeline andmete hoidmine (perfokaardid, perfolindid, magnetlindid)
- 1960ndad aastad – andmebaasid (*DBMS*) → andmete haldus
- 1960ndad, 1970ndad aastad – rakendused (*Online Applications*) → andmevahetus, kaugligipääs
- 1970ndad aastat – personaalarvutid, 4GL personaalsed andmebaasid (1970ndad aastad)
- Relatsioonilised andmebaasid – klient-server lahendused (1980ndad aastad)
- Internet, globaliseeriumine (1990ndad aastad)
  - Järsk andmemahtude kasv
  - Vajadus põhjalikuma analüüsi järele, et saavutada konkurentsieelist
  - Palju erinevaid andmebaase, andmed ei ole seostatavad → vajadus uue lahenduse järele

Andmeladu

W. H. Inmoni järgi: andmeladu on subjektorienteeritud (*subject oriented*), integreeritud (*integrated*), püsiv (*nonvolatile*) ning ajast sõltuv (*time variant*) andmebaas.

- Subjektorienteeritud – st orienteeritud ettevõtte tegevusvaldkondadele või äriprobleemile
- Integreeritud – st ühtlustatakse erinevatest operatiivinfosüsteemidest saadavad andmed
- Püsiv – st andmelaos andmeid ei uuendata
- Ajast sõltuv – andmelao võtmestruktuuriks on aeg

Andmeladu

Operatiivinfosüsteem ...	**Andmeladu ...**
on objektikeskne – objektiks on näiteks klient, teenus jne	on subjektorienteeritud – subjektiks on teenuse osutamine
on tehingupõhine – st korraga tegeldakse vaid ühe objekti (kliendi) andmetega	on analüüsipõhine – st korraga tegeldakse paljude subjektidega (kõik eelmise aasta tehingud)
toetab igapäevaseid tegevusi – st on loodud igapäevaste tegevuste ja toimingute lihtsustamiseks	toetab juhtimisvajadusi – st on abivahendiks ettevõtte strateegiliste juhtimisotsuste langetamisel
on ametnike kasutusalas	on juhtiva personali ja ärianalüütikute kasutusalas
andmed on uuendatavad	olemasolevaid andmeid ei uuendata
on orienteeritud käesolevale hetkele – st andmed kajastavad hetkeseisu	sisaldab väärtusi pikema aja lõikes
operatsioonid korduvad – st teostatavad operatsioonid on päevast päeva samad	heuristiline lähenemine – st ei ole fikseeritud, millist analüüsi nende andmete pealt teostatakse

Andmeladu

- A data warehouse is a **storage architecture** designed to hold data extracted from transaction systems, operational data stores and external sources. The warehouse then combines that data in an aggregate, summary form **suitable for enterprise-wide data analysis and reporting** for predefined business needs.
- The five components of a data warehouse are:
  - production data sources
  - data extraction and conversion
  - the data warehouse database management system
  - data warehouse administration
  - business intelligence (BI) tools

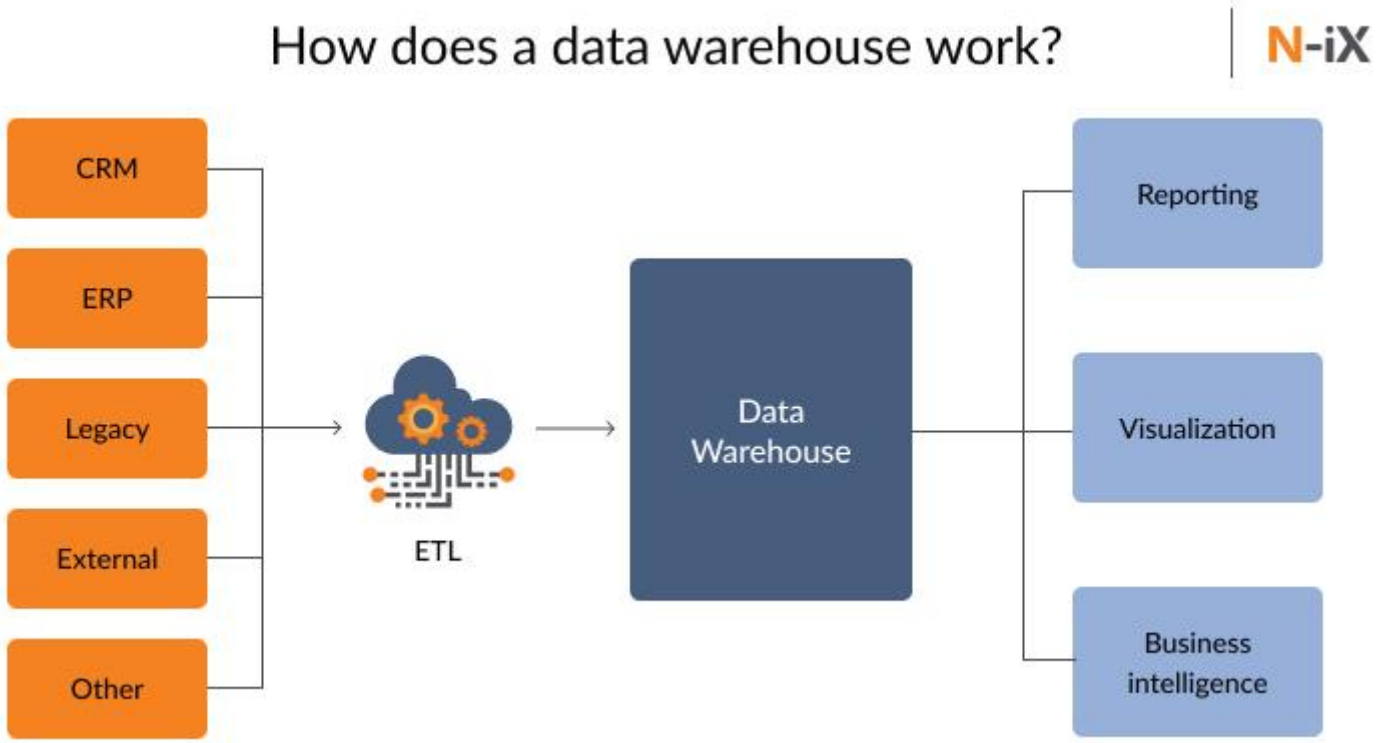
Allikas: Gartner

Andmeladu

Andmeladu on infosüsteem, mis

- koondab erinevatest (operatiiv) infosüsteemidest pärinevad andmed ühtsesse andmehoidlasse
- on orienteeritud aruandlusele ja andmete analüüsimisele (juhtimisotsuste tegemisele), mitte igapäevasele operatiivtööle (teenuste pakkumisele)
- sisaldab järgmisi komponente:
  - andmete laadimise protsessi (sageli koos andmete puhastamisega)
  - andmebaasi/andmeid (summeeritud andmed, detailandmed)
  - aruandlusvahendeid (andmete visualiseerimine, analüütikalahendused)
  - metaandmeid (tehnilised metaandmed, ärilised metaandmed)

Andmeladu


How does a data warehouse work?
N-iX
CRM
ERP
Legacy
External
Other
ETL
Data Warehouse
Reporting
Visualization
Business intelligence

Andmelao kasutamine

- Mis juhtus (*what has happened*)? – Aruandlus (*reporting*), juhtimislauad (*dashboard*), võtmemõõdikud (KPI – *Key Performance Indicator*)
- Miks juhtus (*why it happened*)? – Ad-hoc analüüs, OLAP analüüs
- Mis juhtub tulevikus (what will happen)? – Prognoosimine, ennustav analüüs (*predictive analytics*)
- Mida teha, et juhtuks … (*what to do* …) – What-if analüüs

Ärianalüüsi trendid

- (Peaagu) reaalajas andmete laadimine (*near real-time loading*)
  - Voogandmete kasutamine (*Streaming*)
- Stuktureerimata andmete kasutamine
- Pilvepõhised lahendused
  - Skaleeritavus
- Süvaanalüütika kasutamine
  - Ärianalüütika (*Business Analytics*)
  - Ennustav analüütika (*Predictive analytics*)
  - Andmekaeve (*Data Mining*)
  - Masinõpe (*Machine Learning*)

Mõisteid

- **Ärianalüüs** (*Business Intelligence*) – juhtimisotsuste tegemiseks vajalike andmete kogumiseks, analüüsimiseks ja tulemuste esitlemiseks kasutatavate rakenduste, tehnoloogiate ja oskuste kogum
  - **Business Intelligence** is broad category of applications of applications and technologies for gathering, storing, analyzing, sharing, and providing access to data to help enterprise users make better business decisions
    Allikas: Gartner
  - NB! Eesti keeles kasutatakse mõistet „ärianalüüs“ ka äriprotsesside kaardistamisena
- **Ärianalüütika** (*Business Analytics*) – Analüüsimudelite ja simulatsioonide ehitamine prognoosimiseks ja seoste/sõltuvuste analüüsiks. Siia kuuluvad andmekaeve, ennustav analüüs, statistiliste mudelite ehitamine jms
  - **Business analytics** is comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states. Business analytics includes data mining, predictive analytics, applied analytics and statistics, and is delivered as an application suitable for a business user.
    Allikas: Gartner

Mõisteid

- **Andmeanalüütika** (*Data Analytics*) – toorandmete analüüsimine järelduste tegemiseks. Ärianalüütika on andmeanalüütika erivorm (äriprotsesside kohta järelduste tegemine)
- **Andmekaeve** (*Data Mining*) - andmetest (varjatud) mustrite otsimine
  - NB! Ad-hoc analüüs ei ole andmekaeve
- **Masinõpe** (*Machine Learning*) – arvuti võime „õppida“ e väljastada paremaid tulemeid programmi/algoritmi muutmata
- **Tehisintellekt** (*Artificial Intelligence*) – tehnilise süsteemii võime oma keskkonda tajuda, tajuandmeid töödelda ja ülesandeid lahendada, et saavutada teatav eesmärk

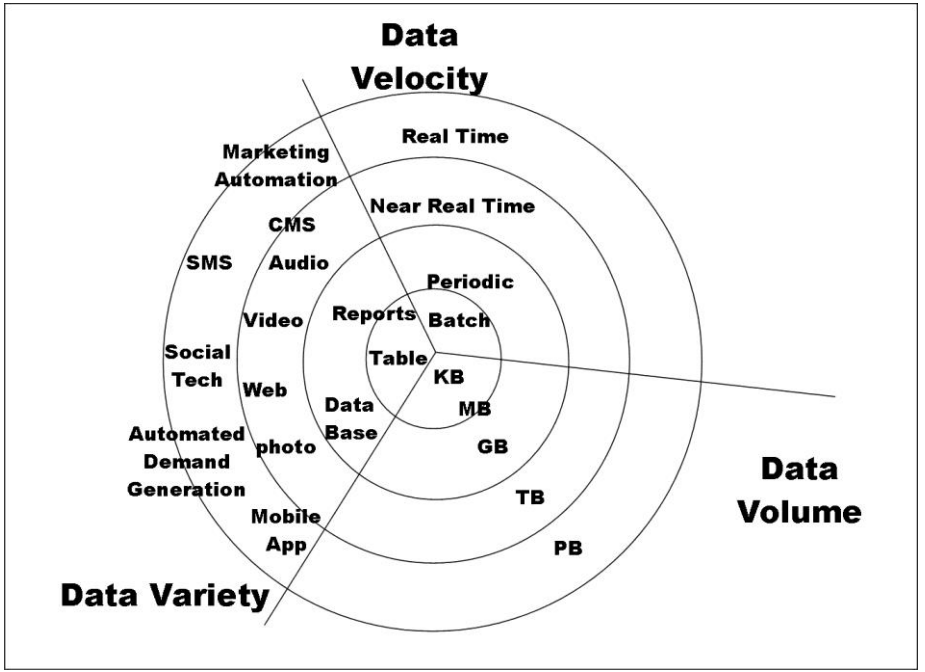
Andmed

- Struktureeritud andmed
  - Andmete koosseis ja struktuur lepitakse kokku enne infosüsteemi loomist
  - (Relatsiooniline) andmebaas – tabelid, veerud/väljad, read/kirjed
- Osaliselt struktureeritud andmed
  - XML, JSON
  - Kasutatakse peamiselt andmevahetuseks
- Struktureerimata andmed
  - Andmete koosseis ja struktuur ei ole kokku lepitud (muutuvad pidevalt)
  - Vabatekst (dokumendid, veebilehed),
  - Pildid, videod
  - E-mailid
  - Sotsiaalmeedia

Suurandmed

- Volume – suur andmete maht
- Variety – andmeformaatide paljusus
- Velocity – andmete tekkimise ja töötlemise „kiirus“ on suur (ja/või varieerub)

- Variability – Andmekogum on ajas muutuv (*inconsistency of data set)*
- Veracity – andmete õigsus „kõigub“



Andmelaoga sarnased lahendused

- Operatiivne andmehoidla (*Operational Data Store* e ODS)
  - Operatiivinfosüsteemide koopia (sageli salvestatud andmelaoga samasse andmebaasi)
- Andmejärv (*Data Lake*)
  - Andmehoidla, kus hoitakse erinevatest allikatest pärinevaid andmeid algsel kujul
  - Sageli seostatakse jagatud failisüsteemidega (Hadoop, Amazon AWS, Microsoft Azure)
  - Andmejärve oluliseks osaks on metaandmed (mis andmed andmejärve on salvestatud). Vastasel juhul on tulemuseks andmesoo (*Data Swamp*)
- Andmejaotur (*Data Hub*)
  - Erinevatest allikatest pärinevate andmete kogum andmete filtreerimiseks ja jaotamiseks
  - Jagab andmeid erinevates formaatides

Andmelaoga sarnased lahendused

	Andmeladu (Data Warehouse)	**Andmejärv (Data Lake)**	**Andmejaotur (Data Hub)**
Eesmärk	Ärianalüüs, ärianalüütika. Ühtse tõe allikas	Andmeteadus, andmekaeve	Andmevahetus (kõik andmed sh operatiivinfosüsteemid)
Andmete formaat	Struktureeritud	Struktureerimata, osaliselt struktureeritud	Enamasti struktureeritud, mitmene andmemudel, vähesel määral struktureerimata
Andmekvaliteet	Kõrge	Keskmine või madal	Kõrge
Andmete haldus (*data governance*)	Kõrge	Madal või puudub	Kõrge
Kasutajad	Juhid, ärianalüütikud	Andmeteadurid	

Paus

09.45–10.00

Andmeait/andmeladu

Andmeladu

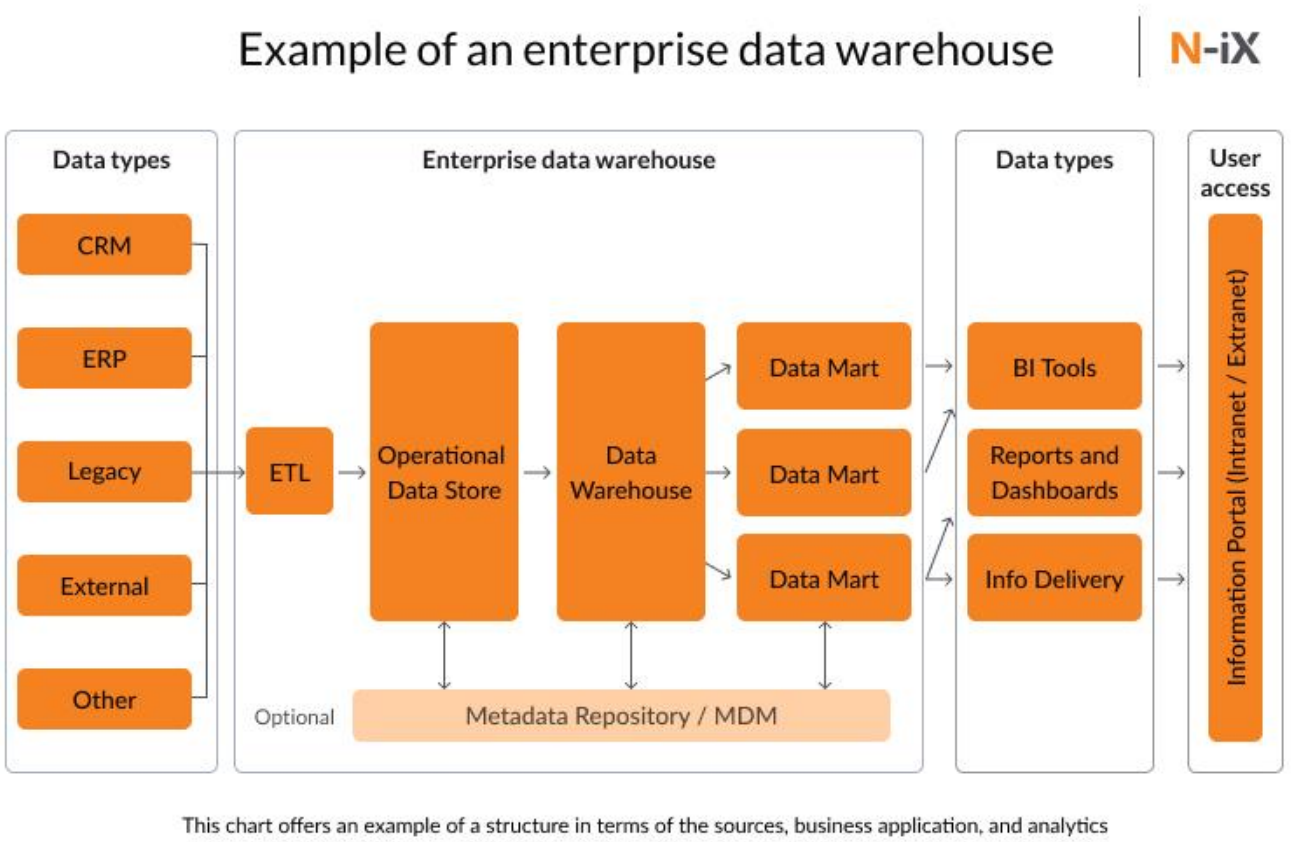
Andmeladu on infosüsteem, mis

- koondab erinevatest (operatiiv) infosüsteemidest pärinevad andmed ühtsesse andmehoidlasse
- on orienteeritud aruandlusele ja andmete analüüsimisele (juhtimisotsuste tegemisele), mitte igapäevasele operatiivtööle (teenuste pakkumisele)
- sisaldab järgmisi komponente:
  - andmete laadimise protsessi (sageli koos andmete puhastamisega)
  - andmebaasi/andmeid (summeeritud andmed, detailandmed)
  - aruandlusvahendeid (andmete visualiseerimine, analüütikalahendused)
  - metaandmeid (tehnilised metaandmed, ärilised metaandmed)

Andmeladu

- Andmealadu ja andmeait on sünonüümid (*Data Warehouse*)
- Andmeladu kitsamas mõttes – andmebaas
- Andmeladu laiemas mõttes – sisaldab lisaks andmebaasile ka andmete laadimist, visualiseerimist ja metaandmeid
- Andmelett e andmevakk (*Data Mart*) – konkreetse osakonna või ärifunktsiooni spetsiifiline osa andmelaost. Võib olla realiseeritud eraldi dimensionaalse mudelina, kuubina, tabelina vms
- Metaandmed (medatata) – andmed andmete kohta
  - Tehnilised metaandmed – andmetüübid, andmete asukohad, seos lähteallikaga
  - Ärilised metaandmed – andmete sisuline tähendus, mõõdikute definitsioonid, äriterminid
- Operatiivne andmehoidla (*Operational Data Store* e ODS) – koopia lähteüsteemi andmetest

Andmeladu



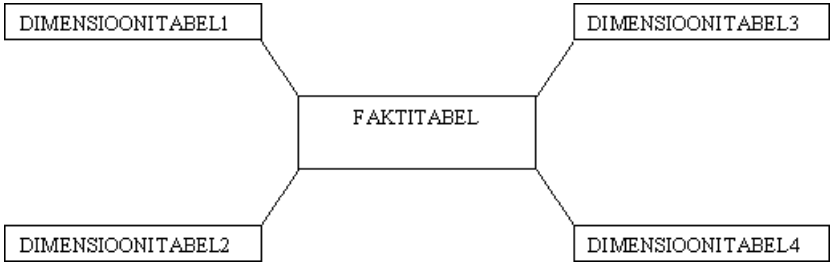
This chart offers an example of a structure in terms of the sources, business application, and analytics

ETL – andmete laadimisprotsess

- ETL – *Extract Transform Load*
- Sisuliselt protsess andmete laadimiseks ja teisendamiseks lähtesüsteemist andmelattu
- Tänapäeval sageli ELT (andmete transformeerimine toimub andmelao andmebaasis)
- Sisaldab
  - Andmete eraldamine lähtesüsteemist (sh muudatuste tuvastamine)
  - Andmete kontrollimine ja ühtlustamine
  - Andmemudeli teisendamine (kolmandalt normaalkujult dimensionaalsele kujule)
  - Andmete laadimine andmelao andmebaasi
  - Teavitused protsessi tulemuslikkusest

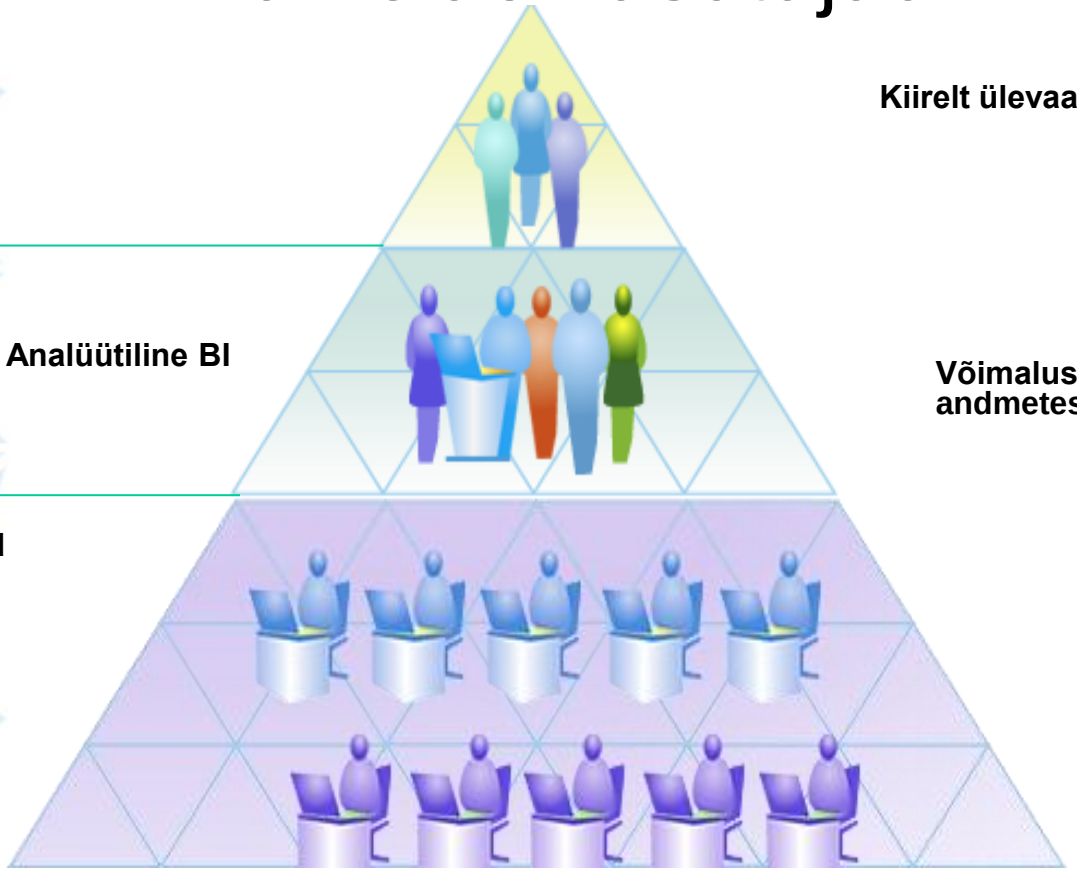
Dimensionaalne andmemudel

- Tähtskeem (*Star Schema*)
- Faktid – ajaga seotud mõõdetav sündmus
  - nt teenuse osutamine
- Dimensioonid – faktiga seotud objektid
  - nt kliendid, teenused jns
- Mõõdikud – atribuut, mille kohta tehakse arvutusi
  - nt tehingute arv, arve summa vms



- NB! Dimensionaalne mudel on denormaliseerutud andmemudel

Andmelao kasutajad



Kiirelt ülevaade olulisest

Analüütikud – Analüütiline BI

Võimaluse „kaevata“ andmetes

Tavakasutajad
Partnerid
Kliendid

-

Taktikaline/ Operatiivne BI

Standardaruandlus

Aruandluse klassifikatsioon

- 0-klõpsu aruanded – aruanded, mis automaatselt saadetakse kasutajale
- 1-klõpsu aruanded e standardaruanded – aruanded, mille tegemiseks tuleb valida aruannete loetelust õige aruanne ja see ühe klõpsuga käivitada
- 2-klõpsu aruanded e parametriseeritud aruanded – aruanded, mille tegemisel kasutajale antakse hulk parameetreid, mille väärtused saab ta enne aruande käivitamist määrata
- 3-klõpsu aruanded e OLAP aruanded – aruanded, mille kuju ei ole ette määratud, vaid selle saab kasutaja määrata ise

- Juhtimislaud (*Dashboard*) – kombinatsiooni võtmenäitajatest (KPI) ja nendega seotud aruannetest
- Lisaks on praktiliselt kõigis aruandlus- ja analüüsitarkvarades vahendid uute aruannete ja juhtimislaudade tegemiseks

Andmekvaliteet

- Andmekvaliteet on andmebaasis leiduvate andmete vastavus reaalsele elule
- Andmekvaliteedist ei saa rääkida, ilma et oleks defineeritud andmekvaliteedi reeglid (st peab olema kirjeldatud, millised on kvaliteetsed andmed)
  - Tehnilised reeglid (väljad täidetud, seosed tabelite vahel)
  - Ärireeglid (väärtuste lubatavus, sündmuste järjekord)
- Andmekvaliteedi analüüs annab ülevaate andmebaasis leiduvatest (potentsiaalsetest) vigadest
- NB! Andmekvaliteedi analüüs vähendab käsitsi kontrollimist vajavate andmete hulka
  - AK analüüsi poolt “vigaseks” tunnistatud andmed ei pruugi olla tegelikult vigased. Tegemist võib olla erijuhtudega. AK analüüsi kontrolli reeglid võivad olla rangemad kui andmete sisestuskontrollide reeglid.

Andmekvaliteet andmelao projektis

- Data warehouses play a crucial role in the success of a business intelligence (BI) program. However, through 2007, more than 50 percent of data warehouse projects will have limited acceptance, or will be outright failures, as a result of a lack of attention to data quality issues, according to Gartner, Inc.

  http://www.gartner.com/press_releases/asset_121817_11.html

- Poor data quality costs US economy around $3,1 trillion a year 2016

  http://www.ibmbigdatahub.com/infographic/four-vs-big-data

Andmekvaliteet andmelao projektis

- Andmekvaliteedi analüüs andmelao projektis võimaldab
  - tutvuda andmetega enne andmemudeli loomist ja ETLi koostamist
  - valida, milliseid andmeid laetakse andmelattu
  - teha kindlaks andmete kustutamise/ülekirjutamise meetodid ning arvestada sellega ETLi realiseerimisel
  - hinnata erinevatest allikatest pärinevate andmete ühildamise keerukust ning valida selle teostamiseks parimad meetodid
  - hinnata andmelaol baseeruva aruandluse ja analüüsi usaldusväärsust
- Andmekvaliteedi hinnangud aruandluses
- Andmekvaliteedi monitooringulahendused

Milleks on vaja andmeladu?

Erinevatest süsteemidest andmete koondamine ühtsesse aruandlus- ja analüüsikeskkonda

- **Ettevõttest tervikpildi saamine** – andmeladu võimaldab seostada omavahel seni eraldiseisnud info (näiteks müügitulemused müügisüsteemist seostatakse töötajate/ müüjate sotsiaalsete näitajatega personali-infosüsteemist)
- **Ühtse tõe printsiip** – eri süsteemid võivad anda samale nimetusele erinevaid tulemusi. Andmelao tegemise (laadimise) üks osa on erinevatest allikatest pärinevate sisuliselt samade andmete ühtlustamine (nt kliendi nimed, aadressid, aga ka toodete/teenuste klassifikatsioonid)
- **Kasutusmugavus** – aruandeid ei pea tegema mitmest erinevast infosüsteemist, mitmed tunnused/mõõdikud/lipud jms on ette arvutatud

Milleks on vaja andmeladu?

- **Jõudlus/kiirus** – andmeladu on optimeeritud analüütilisteks päringuteks (päringud pika ajaperioodi, paljude klientide jms) kohta
  - optimeeritud ja denormaliseeritud andmemudel
  - tehnoloogilised vahendid (OLAP kuubid, vertikaalne andmesalvestus jms)
- **Spetsiaalsed analüüsivahendid**
  - OLAP aruandlus (slice and dice) – tulemuste vaatamine “kõik kõige lõikes”
  - andmete visualiseerimine ja visuaalse analüüsi vahendid
  - prognoosid ja what if analüüs
- **Ajaloo säilitamine**
  - Andmed pikema ajaperioodi kohta
  - Versioneerimine

Sirutuspaus

11.15–11.30

Riigi Infosüsteemi Amet
TARK e-RIIK

Andmeladude tehnoloogiad ja tarkvarad

Riigi Infosüsteemi Amet
TARK e-RIIK

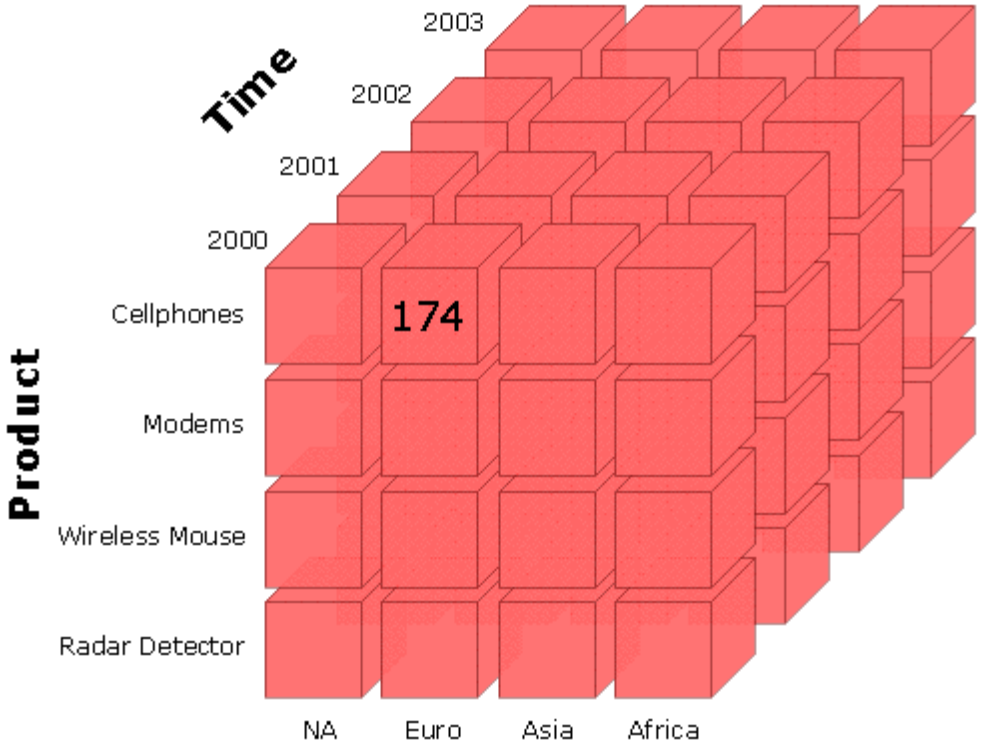
Analüütilised päringud ja nende kiirendamine

- Analüütilised päringud – palju kirjeid, vähe välju (võrdluseks operatiivpäringud – vähe kirjeid, palju välju)
- Analüütilised andmebaasid on optimeeritud analüütiliste päringute käivitamiseks
- Analüütiliste päringute kiirendamise võimalused:
  - OLAP kuubid (*OLAP Cube*)
  - Veerupõhised andmebaasid ja/või andmesalvestus (*Columnar Databases, Columnar storage*)
  - Mäluandmebaasid (*In-Memory Databases*)
  - Paralleeltöötlus (MPP – *Massively Parallel Processing*)
  - Spetsialiseeritud riistvara (*Database appliance*)

OLAP

- OLAP – *Online Analytical Processing* (võrdluseks OLTP – *Online Transaction Processing*)
- OLAP kuup (*OLAP cube*) kui ideoloogia
  - Dimensioonid
  - Hierarhiad on ühes dimensioonis omavahel seotud väljad, milles ühe välja väärtuse valik piirab järgmise välja väärtuste hulka nt aasta → kuu→ päev või ettevõte → tehas → töökoht
  - *slice and dice* – kuubi lõikamine st vaadatakse nt kuubi ühte kihti (nt aruanne kahe dimensiooni lõikes)
- OLAP kuup kui tehnoloogia (MOLAP)
  - Sisaldab ettearvutatud summasid kõigi dimensioonide kõigi liikmete lõikes
  - Igal tarkvaratootjal erinev tehnoloogia ja erinev päringukeel (Microsoft - MDX, Oracle jne)

OLAP kuup


Time
2003
2002
2001
2000
Product
Cellphones
Modems
Wireless Mouse
Radar Detector
174
NA
Euro
Asia
Africa
Location

Veerupõhine andmesalvestus

- Andmeid hoitakse veerupõhiselt
- Eelised
  - Pakkimine – Sarnased andmed kõrvuti, suurem pakkimistihedus
  - Analüütiliste päringute kiirus kuni 100x suurem, kui reapõhisel andmesalvestusel
  - Puudub vajadus summeeritud andmestike järele
- Veerupõhised andmebaasid (nt Vertica)
  - Klassikaline ANSI-SQL andmebaas (ei vaja uue keele õppimist)
  - Rohkem erinevaid indekseid/projektsioone (samu andmeid võimalik optimeerida erinevate päringute jaoks)

Table

	Country	Product	Sales
Row 1	India	Chocolate	1000
Row 2	India	Ice-cream	2000
Row 3	Germany	Chocolate	4000
Row 4	US	Noodle	500

Row Store

Row 1	India
	Chocolate
	1000
Row 2	India
	Ice-cream
	2000
Row 3	Germany
	Chocolate
	4000
Row 4	US
	Noodle
	500

Column Store

Country	India
	India
	Germany
	US
Product	Chocolate
	Ice-cream
	Chocolate
	Noodle
Sales	1000
	2000
	4000
	500

Mälu-andmebaasid

- Kogu andmebaas paikneb operatiivmälus
- Sobib väiksemate andmemahtude korral ja spetsiifiliste analüütikaülesannete lahendamiseks
- Ei pruugi sobida real-time lahenduste jaoks (vajab andmete laadimist)
- Mitmed ärianalüüsi tarkvarad (nt PowerBI) kasutavad mälu-andmebaasi aruannete näitamiseks (loeb andmed mällu ja lubab neid seal töödelda nt *slice-and-dice)*

ETL tarkvarad

- Kasutamine sõltub vajadustest
  - Andmebaaside liidesed
    - *Bulk load* andmelao andmebaasi
  - Teenuste kasutamise võimekus (nt REST, X-tee)
  - Erinevate andmeformaatide töötlemine (XML,JSON, Excel jne)
  - Paralleltöötlus
  - Voogandmete töötlus
- Integreeritud andmebaasiga (nt Microsoft, Oracle, SAP) või eraldiseisev (Pentaho)

Apache Stack

- Mõeldud eelkõige suurandmete salvestamiseks ja töötlemiseks
- Sageli spetsiifilised keeled (MapReduce, Pig, Spark jne)
- Apache komponente
  - Hadoop
    - Jagatud failisüsteem – *Distributed FileSystem* (HDFS)
    - Arvutuste raamistik – *Computation or Processing framework* (MapReduce).
  - Hive – SQL sarnane liides Hadoop peal
    - Teisendab SQL päringud MapReduce päringuteks
  - HBase – mitterelatsiooniline jagatud andmebaas
  - Pig – Programmeerimiskeel Hadoop platvormile
    - Andmete analüüsimiseks ja ETLi (andmete transformeerimise) koostamiseks
  - Spark – programmeerimiskeel suuremahuliste andmte töötlemiseks
  - Kafka – voogandmete edastamine/töötlemine
  - Airflow – töövoogude haldamise keskkond
    - ETLi protsesside ajastamine

Jagatud andmetöötlus

- Mitmed andmelao lahendused/tarkvarad toetavad jagatud/klaster lahendusi
  - Andmed ja/või andmete töötlemine jagatakse mitme erineva serveri vahel
  - Skaleeritavus – süsteemi jõudluse tõstmiseks võib lisada täiendavaid servereid
  - Tõrkekindlus – üksiku serveri mahakukkumine ei tähenda terve süsteemi töö katkestust

- Vertica – multi node
- Apache Stack – kõik vahendid disainitud jagatud andmesalvestuseks ja töötluseks

Süvaanalüütika

- Andmebaasis paiknevad analüüsivahendid
  - Nt Vertica – aegridade analüüs, regressioonanalüüs, K-keskmised, tugivektor-masin jne
- Visualiseerimisvahendites (Tableau jt) olevad võimalused
  - Ennustav analüüs (regressioonmudelid)
  - Võimalus liidestada R ja/või Pythoniga
- Andmete analüüsimine R-i või Python'iga

Lõunapaus

12.00–13.00

Andmelao projekt

Riigi Infosüsteemi Amet
TARK e-RIIK

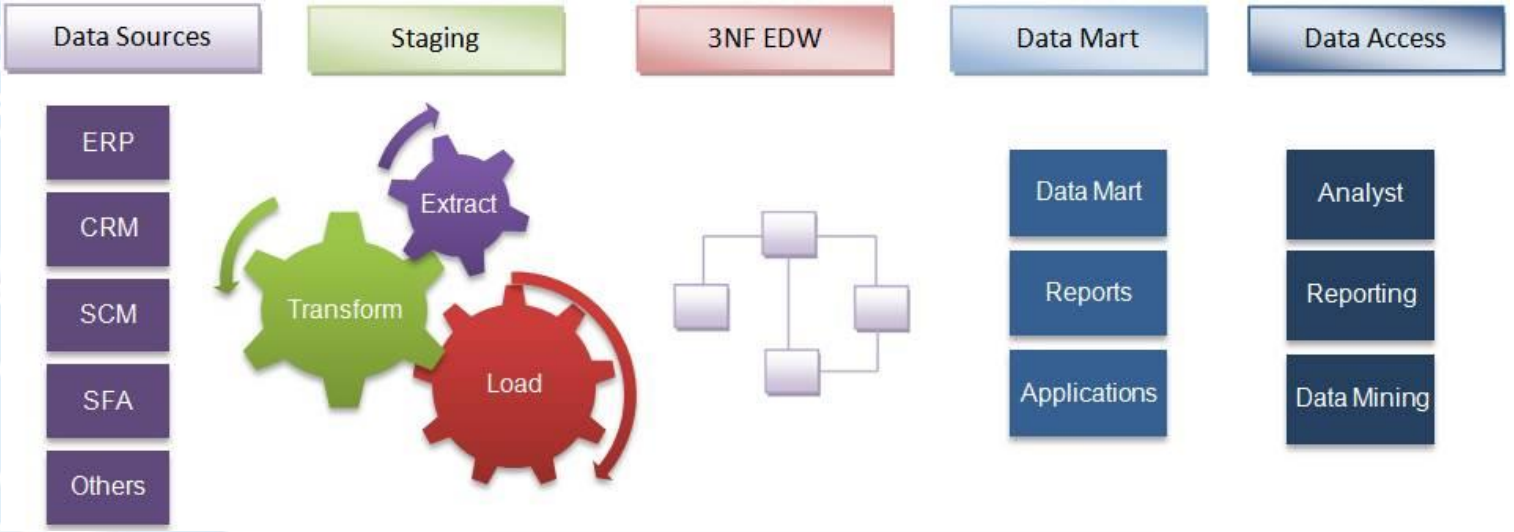
Andmelao ehitamise metoodikad

- Top-down design (Bill Inmon)
  - alustatakse ettevõtte ülese andmelao (EDW) loomisest
- Bottom-up design (Ralph Kimball)
  - alustatakse andmelettide loomisest, andmeletid moodustavad andmelao
- Hybrid design
  - andmelettide ja ettevõtte ülese andmelao loomine toimub paralleelselt (etappide kaupa, vaheldumisi)
- Federated design
  - ei ole päriselt metoodika vaid põhimõte „share the “highest value” metrics, dimensions, and measures wherever possible, however possible“

Allikas: http://tdan.com/four-ways-to-build-a-data-warehouse/4770

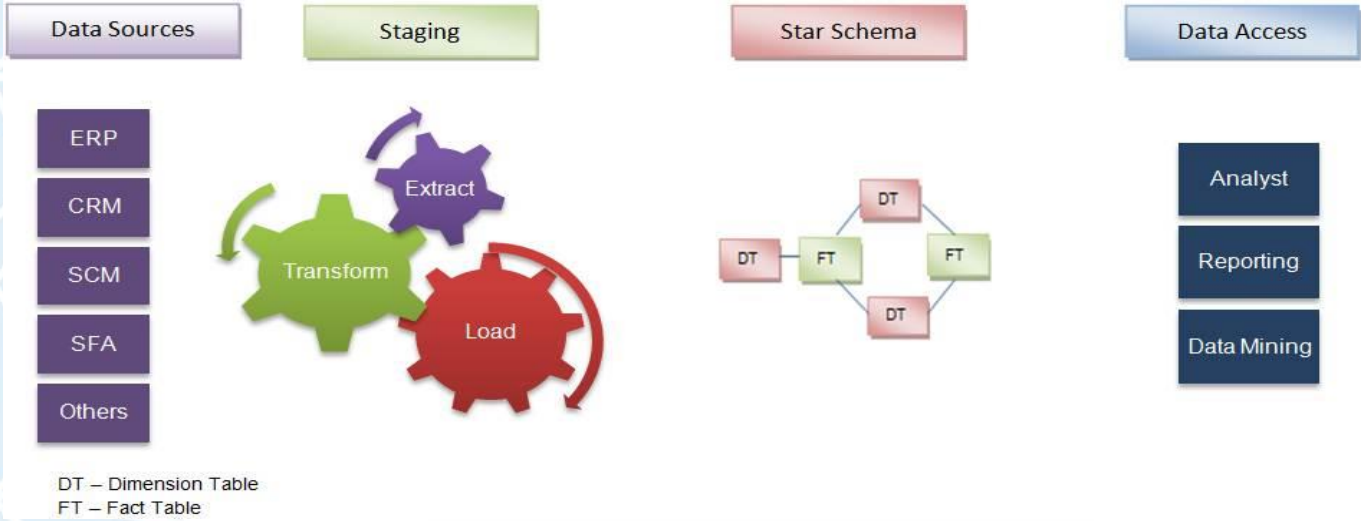
AL metoodikad – *top-down design*

- Kõigepealt luuakse ettevõtte kõigi infosüsteemide ülene, detailandmeid sisaldav, kolmandal normaalkujul andmebaas – Enterprise DW
- Andmeletid luuakse hiljem (pärast EDW valmimist)



AL metoodikad – *bottom-up design*

- Kõigepealt luuakse valdkonnapõhised andmeletid kasutades dimensionaalset andmemudelit
- Loodud andmeletid kombineeritakse ettevõtte-üleseks andmelaoks
- NB! Sellise disainiga andmeladu (tüüpiliselt) ei sisalda detailandmeid



AL metoodikad – *hybrid ja federated design*

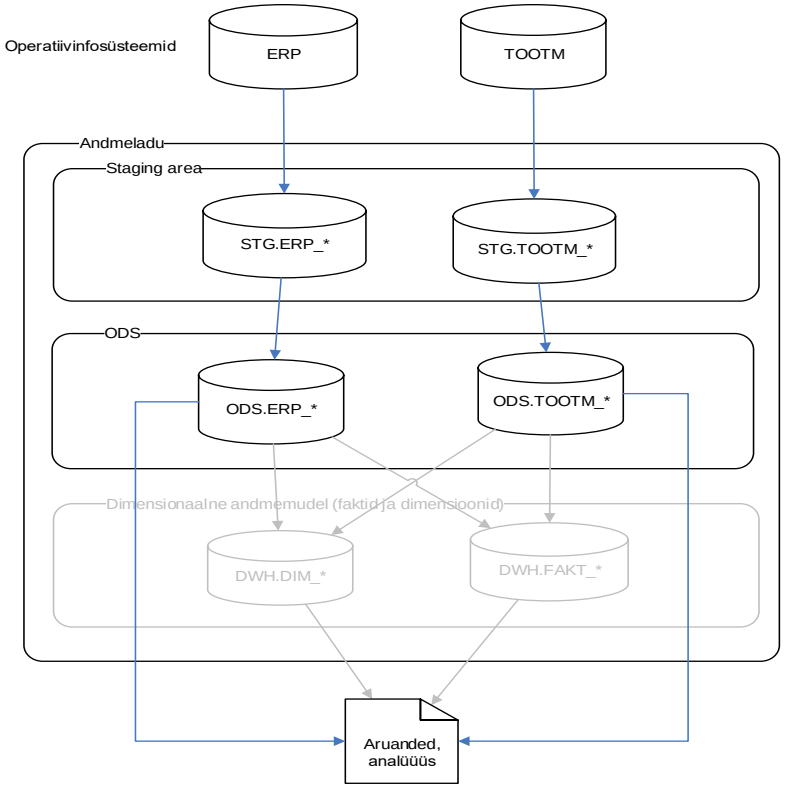
- Hybrid design
  - Kombinatsioon „bottom-up“ ja „top-down“ metoodikatest
  - EDW ja andmeletid luuakse paralleelselt (ca 2 nädalat EDW mudeldamine, et saada üldine ülevaade; siis esimesed andmeletid lõpuni, et saada kasutatav lahendus; siis vastavalt vajadusele EDW täitmine andmetega)
- Federated design
  - ei ole päriselt metoodika, vaid põhimõte „share the “highest value” metrics, dimensions, and measures wherever possible, however possible“

Operational Data Store + andmeladu

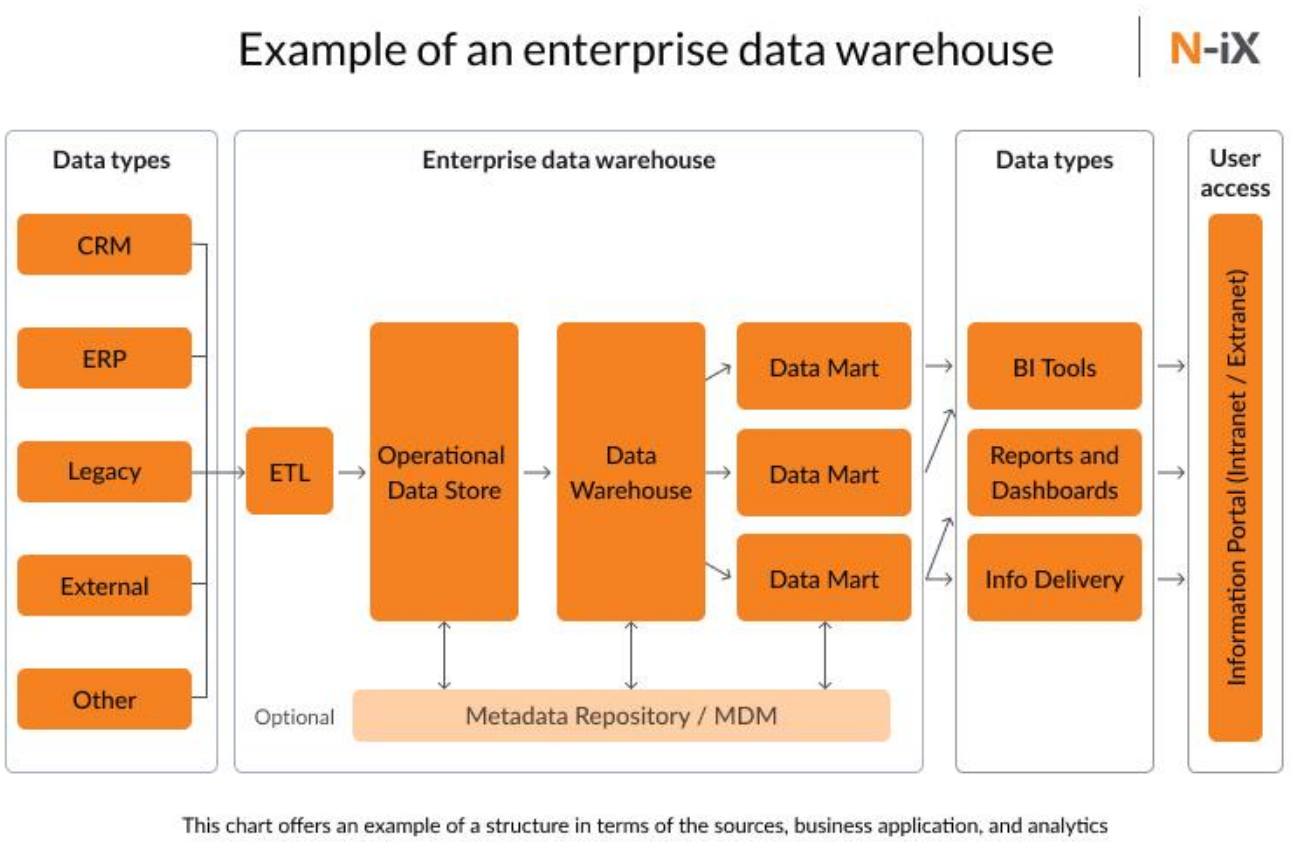
Jaotada andmeladu kaheks osaks – lähteallikate põhine ODS (*operational data store*) + dimensionaalne DW (*data warehouse*)

Eelised

- ODS laadimine on võimalik teha allika kaupa → ETLi lahendus on standardne ja kiirelt valmiv
- Esimesed aruanded võib teha ODS andmete pealt (koopia operatiivsüsteemi aruannetest) → esimesed tulemid valmivad kiirelt
- ODS laadimine on võimalik teha peaaegu-reaalajas (transformatsioone sisuliselt ei ole) → Operatiivaruanded reaalajas ODS pealt
- Dimensionaalse DW (ühtse tõe printsiip) laadimine intervalliga, analüütilised päringud DW pealt
- ODS sisaldab kõiki detailandmeid → kõik analüüsid on võimalikud



Andmeladu



This chart offers an example of a structure in terms of the sources, business application, and analytics

Metoodika

- Parallelselt
  - ODS laadimise realiseerimine (allikate kaupa)
    - Ligipääsud
    - Muudatuste/kustutamiste tuvastamine
    - ODS laadimise realiseerimine
    - Alglaadimine
    - Reg. laadimiste kävitamine
  - Dimensionaalse mudeli loomine (andmelettide kaupa)
    - Ärivajaduste kaardistamine
    - Dimensionaalse mudeli/andmelettide disain
- Dimensionaalse mudeli/andmelettide laadimise realiseerimine
- Aruannete/juhtimislaudade loomine

Resta ETL lahendus

- Automaatne/dünaamiline andmelao andmemudeli (STG ja ODS kiht) loomine/muutmine vastavalt lähtesüsteemi andmemudelile ja selle muudatustele
  - Lähtesüsteemi metaandmete import
  - Lähtesüsteemi (hetkel kehtiva) andmemudeli ja andmelaole teadaoleva andmemudeli võrdlus ja muudatuste tuvastamine
  - Andmelao andmemudeli muutmine/täiendamine
- STG ja ODS kihtide automaatne laadimine koodi kirjutamata
- Erinevad meetodid ODS kihi tabelite laadimiseks:
  - LOAD_MODIFIED – muutunud andmete laadimine
  - REWRITE – andmete ülekirjutamine
  - FULL_COMPARE – andmete võrdlemine
  - OBJECT_COMPARE – objekti tasemel võrdlus
- Vajadusel andmete pseudonüümimine

Resta ETL lahendus

- Raamistik API liidese kaudu andmete laadimiseks (spetsiifilised komponendid tuleb kirjutada vastavalt konkreetsele API liidesele)
- Dimensionaalse mudeli/andmelettide laadmine (vastav kood eraldi kirjutada)
  - tabelite vaheliste sõltuvuste arvestamine – kui alustabel saab vea, siis tabelit/andmeletti ei laadita
- Logimine, veahaldus ja laadimise tulemuslikkusest (vigadest, muudatustest) teavitamine (näiteks e-postiga)
- Realiseeritud Pentaho + Vertica platvormil
- Võimalikult parametriseeritud, st mitmeid laadimise aspekte saab muuta koodi muutmata (st tarnet tegemata)

Projekti võtmerollid - tellija

- Projekti sponsor/tellija (*Project Sponsor*)
  - Tavaliselt ettevõtte juhtkonnast
  - Tellija esindaja, ressursside eraldamine
- Projektijuht (*Project Manager*)
  - Projektiplaani ja arendusskoobi täitmise jälgimine, ülevaated sponsorile
- Valdkonna spetsialist (*Subject Matter Expert*)
  - Lõppkasutaja esindaja, teab äriprotsesse ja infovajadusi
  - Vastuvõtutestid
- Haldur/Administraator (*DBA*)
  - Andmehaldur – loogiline haldamine (nt andmete laadimine)
  - Administraator – tehniline haldamine (serverid, backup)

Projekti võtmerollid – täitja

- Projektijuht (*Project Manager*)
  - Projektiplaani ja arendusskoobi täitmise jälgimine, arendusmeeskonna tööde korordineerimine, ülevaated tellijale
- Analüütik (*Functional Analyst*)
  - Ärivajaduste kaardistamine (meetrika, aruanded, juhtimislauad jms)
- Ärianalüüsi arhitekt (*BI Architect*)
  - Lahenduse üldine disain
  - Andmemudeli koostamine
- ETLi arendajad (*ETL Developers*)
  - ETL lahenduse ja analüütikalahenduse realiseerimine
- Analüütikalahenduse arendajad (*Report Developers*)
  - Aruannete, juhtimislaudade jms realiseerimine

Andmelao projekti riskid

- **Üks suur fikseeritud eelarvega projekt**
- Projekti eesmärk, sisendid ja/või väljundid ei ole piisavalt defineeritud
  - Nt on nimetatud lähtesüsteem aga pole kirjeldatud, mis andmeid on vaja või mis andmetega tehakse
- Projekti käigus muutuvad eesmärgid ja/või vajadused
- Puuduvad vajalikud andmed või andmed ei ole kasutataval kujul
  - Nt andmed PDF formaadis
- Puudub andmelao projekti kogemus ja/või lähtutakse operatiivinfosüsteemi projektist
  - Nt "Lahendus realiseerida objektorienteeritult"
  - Testimine ainult testandmetega
- Puudub projekti sponsor või ei ole temalt piisavalt toetust projektile

Andmelao projekti riskid

- Ebapiisav/ebapädev meeskond (nii tellija kui täitja poolel)
  - sh hõivatus teiste tööde/projektidega
- Andmekvaliteet
- Teiste seotud projektide (sh lähtesüsteemide arendajad, välised kasutajad, seaduse muudatused jne) mõju projektile

Andmelao haldus

- Laadimiste jälgimine, tõrgetele reageerimine
  - Ühenduste vead
  - Andmevead
  - Lähtesüsteemide arendused
    - Õigel ajal teavitamine
    - Automaatsed muudatuste tuvastamised
- Uued arendused
  - Asutuse ärivajaduste muutused
  - Uued infosüsteemid
  - Seadusemuudatused

Riskide maandamine - parimad kogemused

- Agiilseid arendusmeetodid (Scrum)
  - 2 nädalased sprindid
  - Pidev tööde ülevaatamine, hindamine ja prioritiseerimine (*grooming, retrospective*)
  - Iga hommik *stand-up*
  - Iga kuu tarne
- Lahtine ja piisav eelarve – 2-4 täiskohaga arendajat
- Piisav/tugev meeskond tellija poolel
  - Analüütikud (valdkonna teadmine)
  - IT tehniline meeskond (tehniline tugi)
- Koostöö ja suhtlus – kiired vastused küsimustele
- Koolitused
- Andmelaol oluline/selge koht ettevõtte strateegias

Paus

14.00–14.15

Andmeladu ja Eestis kehtivad nõuded

Riigi Infosüsteemi Amet
TARK e-RIIK

Andmete dubleerimine andmelaos

- Avaliku teabe seaduse § 43^3 lõige 2: "Keelatud on asutada ühtede ja samade andmete kogumiseks eraldi andmekogusid."
- Kui andmeid vaid töödeldakse (nagu andmeaitades, kus andmed moodustatakse teiste andmekogude baasil ning neid spetsiaalselt ei koguta), siis on dubleerimine lubatud

Allikas: V. Kotkas, H-M. Haav, J. Tepandi, E. Õunapuu, J. Grauberg, Uurimisprojekti "Andmeaitade (teiseste andmekogude) loomise põhimõtete väljatöötamine" lõpparuanne, TTÜ Küberneetika Instituut, 2013

Eesti ristfunktsionaalsuse nõuded

- https://koodivaramu.eesti.ee/e-gov/cfr/-/blob/master/cfr.xml
- Osa nõudeid on võimalik täita vaid siis, kui need on täidetud lähtesüsteemis
  - nt "Objektid identifitseerida registrikoodide abil", "Rakendada aadressiandmete süsteemi nõudeid" jne
- Osa nõuded kirjutada selgelt välja hankedokumentidesse sh **arvestada vastava eelarvega**
  - nt "Rakendus peab olema läbinud enne toodangusse minemist turvatestimise", "Toodangusse ei lähe kood, mis pole läbinud koodi ülevaatamist (code review)"
- Eesti ristfunktsionaalsuse nõue 22: "Infosüsteemide vaheline andmevahetus toimub üle X-tee" (järgmisel slaidil)

X-tee

- Eesti ristfunktsionaalsuse nõue 22: "Infosüsteemide vaheline andmevahetus toimub üle X-tee"
- Andmevahetus X-tee kaudu andmelaos
  - X-tee infrastruktuur on hea, kuid sageli puuduvad sobivad teenused
  - Enamik teenused on mõeldud üksiku objekti andmete pärimisele
  - Enamasti puuduvad teenused andmelao alglaadimise teostamiseks
    - Ühel juhul võttis andmeallika alglaadimise tegemine üle kuu aja (mitmed tõrked sh serverite logiruum täis, teenused ei vasta jms)
  - Uus RIHA
    - vajalike/sobilike teenuste leidmine raske
    - sageli puudub ajakohane teenuste dokumentatsioon
  - Sõltuvus kolmandast osapoolest teenuste tegemisel/täiendamisel
    - Nt oodatakse mitmendat aastat, millal vajalik teenus valmib
  - Soovitan kasutada ainult asutuse välistest allikatest andmete laadimisel (mitte asutuse sisestest infosüsteemidest laadimisel)

Detailandmete/isikuandmete säilitamine andmelaos

- Milleks andmeladu kasutatakse
  - Kui kasutatakse ka operatiivaruandluseks – siis on vajalik hoida detailandmeid/isikuandmeid
  - Kui kasutatakse ainult summeeritud või statistlilseks aruandluseks, siis võib detailandmeid/iskuandmeid andmelattu mitte laadida
- Kust lõpeb lähteinfosüsteem ja algab andmeladu
  - Lähteinfosüsteemi põhised skeemid (STG, ODS, DIM). Neid kasutada operatiivaruandluseks.
  - Eraldi skeem nendele andmetele, mis baseeruvad mitmel allikal. Siin rakendada andmete pseudonüümimist/anonüümimist

Detailandmete/isikuandmete säilitamine andmelaos

- Andmete pseudonüümimine
  - Isikuandmete töötlemine sellisel viisil, et isikuandmeid ei saa enam täiendavat teavet kasutamata seostada konkreetse andmesubjektiga (äratuntavate isikuandmete asendamist varjunimede, numbrikoodide ja muude tunnustega, mida asjassepuutumatud isikud ei oska ära arvata)
  - Vajaduse korral on andmed võimalik algsele kujule viia
  - Näiteks AES krüpteerimine
  - Kasutatakse juhul kui on vaja säilitada ühe isiku/subjekti tegevuste omavaheline seostamine (nt teenuste tarbimine üle kogu ajaloo)
- Andmete anonüümimine
  - teabest kõikide jälgede kaotamist, mis võiksid viia tuvastatavate isikuteni (kustutamine, agregeerimine)
  - ei võimalda algse kuju taastamist
  - anonüümimist võib kasutada rakenduse tasemel (aruanne ei väljasta detailandmed või summeerutud andmeid, kui elemente on vähem kui fikseeritud arv)

Andmete pseudonüümimise näide

- Lähteallikate põhised skeemid, õigused sarnaselt lähteallika kasutajatega
- Andmeladu sisaldab 2 "andmebaasi"
  - Aktiivbaas – Isikustatud andmed, maksimaalselt 10 aasta andmed
  - Pseudobaas – Kogu ajalugu aga pseudonüümitud kujul (osadest andmeväljadest kustutatakse andmed, osades väljades asendatakse andmed pseudo koodiga, osades väljades muudetakse andmete täpsusastet nt aadress valla täpsusega)
- Välistele allikatele (nt statistiliste mudelite ehitamine) edastatakse ainult minimaalne hulk ja pseudonüümitud andmeid
- Testkeskkonna "sogastamine"
  - kord kvartalis tehakse kõikidest süsteemidest live baasi koopia testi
  - koopia tegemise käigus isikuandmed "sogastatakse" – isikuandmed anonüümitakse, isikud segatakse nii, et ühe isiku teenuste tarbimine liigub teise isiku külge
  - Iga "sogastamise" järel teostatakse andmelao täislaadimine

Küsimused/vastused

Aitäh!

Koolituse korraldas Majandus- ja Kommunikatsiooniministeerium Euroopa Liidu struktuuritoetuse toetusskeemist "Infoühiskonna teadlikkuse tõstmine", mida rahastab Euroopa Regionaalarengu Fond.

Koolituse aitas läbi viia Conference Expert / Reisiekspert

RIIGI INFOSÜSTEEMI AMET

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

TARK

e-RIIK