

TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

Agne Rääbis
Analüütikaprojektide realiseerimise
tehnoloogilised lahendused
Magistritöö (15 EAP)

Juhendajad: Anne Villems, MSc
Ain Rääbis, MSc

Analüütikaprojektide realiseerimise tehnoloogilised lahendused

Lühikokkuvõte:

Käesolev magistritöö analüüsib ja kirjeldab peamisi tehnoloogiaid, mida kasutatakse suuremate analüütikaprojektide korral: andmeladu, andmejärve, loogilist andmeladu ja andmejaoturit. Analüüs sisaldab eestikeelses teaduskirjanduses seni käsitlemata tehnoloogilisi lahendusi. Töö sisaldab ülevaadet tehnoloogiate arhitektuurist, tugevustest ning nõrkustest. Lisaks võrdleb magistritöö käsitletud tehnoloogiaid ning koondab tehnoloogiate omadused tabelina.

Magistritöö lõpptulemust saavad praktilise sisendina kasutada analüütikavaldkonna projektide tellijad ja arendajad. Töö annab info uuematest valdkonna tehnoloogiatest ning vajaliku sisendi paremate lahenduste pakkumiseks. Lisaks saab magistritööst sisendi, mis aitab valida ettevõtte vajadustest lähtuva tehnoloogilise lahenduse.

Võtmesõnad: Andmeladu, andmejärv, loogiline andmeladu, andmejaotur, ärianalüüs

CERCS: P175 Informaatika, süsteemiteooria

Technological Solution's in Data Analytics Projects

Abstract:

This master's thesis analyzes and describes the main technological solutions used in large analytical projects: data warehouse, data lake, logical data warehouse and data hub. Analyzes includes technological solutions that have not been described in the Estonian scientific literature before. The master's thesis provides an overview of the technologies architecture, advantages and disadvantages. In addition, the characteristics of the technologies are compared.

The master's thesis can be used as a practical input by clients and developers in the field of analysis. The work provides information about the newest technologies and provides developers the necessary input for offering their clients the best solution. Work's result can also be used as assistant tool for choosing the right technological solution based on the company's needs.

Keywords: Data warehouse, data lake, logical data warehouse, data hub, business intelligence

CERCS: P175 Informatics, systems theory

Sisukord

Sissejuhatus.....	4
Lühendid ja mõisted	6
1. Analüütika valdkonna tutvustus	8
1.1 Ärianalüüsi ning ärianalüütika mõiste	8
1.2 Ärianalüüsi eesmärk.....	8
1.3 Ärianalüüsi võimalused.....	9
1.4 Ärianalüüsis kasutatavate andmete tüübid	10
1.5 Ärianalüüsi projektide tehnilised lahendused.....	10
1.6 Ärianalüüsi trendid	10
2. Analüütikaprojektide realiseerimise tehnoloogilised lahendused	12
2.1 Andmeladu	12
2.1.1 Andmelao arhitektuur.....	15
2.1.2 Andmelao tugevused	20
2.1.3 Andmelao nõrkused	20
2.1.4 Andmelao kokkuvõte	21
2.2 Andmejärv	21
2.2.1 Andmejärve arhitektuur.....	23
2.2.2 Andmejärve tugevused	26
2.2.3 Andmejärve nõrkused	27
2.2.4 Andmejärve kokkuvõte	28
2.3 Loogiline andmeladu	29
2.3.1 Loogilise andmelao arhitektuur	30
2.3.2 Loogilise andmelao tugevused.....	32
2.3.3 Loogilise andmelao nõrkused	33
2.3.4 Loogilise andmelao kokkuvõte	33
2.4 Andmejaotur	34
3. Käsitletud tehnoloogiate võrdlus	37
Kokkuvõte	40
Viidatud kirjandus	42
Lisa 1. Litsents	45

Sissejuhatus

Ettevõtte edukus sõltub suurel määral sellest kui hästi suudab ettevõtte oma tegevuseks vajalikke andmeid koondada, talletada ning enda kasuks tööle panna. Lahenduseks on siinkohal spetsiaalsed analüütikaprojektide realiseerimise tehnoloogiad nagu näiteks andmeladu. Jätkusuutliku ning ettevõtte tulevikuperspektiive arvestava tehnoloogia valiku langetamiseks on aga vaja esmalt mõista erinevaid tehnoloogiaid.

Töö kirjutamisel püstitati hüpotees, et eestikeelses teaduskirjanduses ei ole varem vaadeldud analüütikalahendusi laiemas vaates, kus käsitletakse erinevaid (sh ka kõige uuemaid tehnoloogiaid nagu näiteks andmejärv ja loogiline andmeladu) tehnoloogiaid koos nende võrdlusega. Selline käsitus annaks kokkuvõtva ülevaate ettevõttele, kes tahab endale ehitada tänapäevast analüütikalahendust ning kellel on vaja valida tema ettevõtte vajadustest lähtuv sobiv tehnoloogiline lahendus.

Käesoleva magistritöö eesmärk on kontrollida püstitatud hüpoteesi, analüüsides olemasolevaid kirjalikke allikaid, mis käsitlevad analüütikaprojektide realiseerimise tehnoloogiaid. Lisaks on magistritöö eesmärgiks luua eestikeelne allikas, mis kirjeldab uuemaid analüütikalahenduste tehnoloogiaid.

Magistritöös on püstitatud järgmised uurimisülesanded:

- 1) uurida, millised on peamised analüütikaprojektide realiseerimise tehnoloogiad, mida kasutatakse suuremate projektide korral;
- 2) anda ülevaade uurimise käigus leitud tehnoloogiatest;
- 3) analüüsida ning kirjeldada uurimise käigus leitud tehnoloogiate omavahelisi erinevusi.

Magistritöös analüüsitakse ja kirjeldatakse analüütikaprojektide realiseerimise tehnoloogiaid suurte analüütikaprojektide vaatest. Suur lahendus on selline, mis kasutab mitmeid erinevaid andmeallikaid, erinevaid andmeformaate, erinevaid andmete ligipääsemise meetodeid ning lahendusel on mitmeid erineva profiiliga kasutajaid. See tähendab, et suurel analüütikalahendusel on väga palju erinevaid nõudeid ja vajadusi. Samas leidub nende nõuete täitmiseks mitmeid erinevaid tehnoloogiaid.

Tehnoloogiate analüüsil on magistritöös lähtutud järgmistest küsimustest:

- 1) millised on lahenduses kasutatavad andmed;

- 2) kes ja milleks neid andmeid kasutab;
- 3) kuidas andmeid talletatakse;
- 4) kuidas liigutatakse andmeid analüütikalahendusse ja selle sees.

Magistritöö jaguneb kolmeks osaks. Esimeses osas tehakse sissejuhatus analüütikavaldkonda ning selgitatakse peamisi mõisteid. Töö teises osas tutvustatakse analüütikaprojektide realiseerimise tehnoloogiaid – nende olemust, arhitektuuri, tugevusi ja nõrkusi. Töö kolmandas osas võrreldakse käsitletud tehnoloogiaid ning esitatakse tehnoloogiate omadused koondtabelina.

Lühendid ja mõisted

Andmejärv (*Data Lake*) – tehnoloogiline lahendus, mis võimaldab koguda, säilitada, töödelda ja tagada kasutajatele ligipääs mistahes (st nii struktureeritud kui ka struktureerimata) andmetele.

Andmekaeve (*Data Mining*) – protsess, mille käigus analüüsitakse andmete peidetud mustreid, et muuta need andmed kasulikuks ja kasutatavaks teabeks. Andmed andmekaeve jaoks kogutakse ja koondatakse erinevatest allikatest, näiteks andmeladudest. Andmekaeve algoritmid otsivad andmetest mustreid, mida ettevõtte saaksid kasutada paremate otsuste tegemiseks. Näiteks saab andmekaeve abil teha paremaid otsused, mis aitavad kulusid vähendada, tulusid suurendada või klientidele paremat teenust pakkuda.

Andmelett (*Data Mart*) – ühe konkreetse äriülesande lahendamiseks mõeldud andmelao osa. Andmelett annab kasutajale juurdepääsu konkreetset tüüpi andmetele, mille kasutamise järele esineb sage vajadus.

Andmemudel (*Data Model*) – andmebaasi mudel, mis defineerib ära loogika, millised andmed ning millistel tingimustel andmebaasi salvestatakse. Andmemudel määrab ära andmete koosseisu ja struktuuri. Andmemudelit kujutatakse enamasti graafilisel kujul ning see on koostatud äriprotsessist lähtuvalt.

API (*Application Programming Interface*) – rakendusliides ehk programmiliides. Reeglistike kogum, mida kasutatakse erinevate tarkvarateenuste omavaheliseks suhtluseks nagu näiteks suhtlus andmevahetuse teenuse osas [1].

Asjade internet (*Internet of Things*, lühend IoT) – osa internetist, mille moodustavad võrku ühendatud ning omavahel andmeid vahetavad nutiseadmed. Asjade interneti alla kuuluvad näiteks võrku ühendatud sensorid ning nende poolt mõõdetud andmed.

Ennustav analüütika (*Predictive analytics*) – täiustatud analüütika vorm, mis uurib andmeid eesmärgiga saada vastused küsimustele nagu „Mis hakkab juhtuma?” või „Mis tõenäoliselt juhtub?” [2].

ETL (*Extract, Transform, Load*) – protsess, mis vastutab andmete andmebaasi sisestamise, töötlemise ning laadimise eest.

ELT (*Extract, Load, Transform*) – protsess, mis vastutab andmete andmebaasi sisestamise, laadimise ning töötlemise eest. Sisuliselt on ETL-i ja ELT näol tegemist sama lahendusega, erineb vaid vastavate alamprotsesside tööjärjekord.

Juhtimislaud (*Dashboard*) – ehk juhtimise töölaud. Juhtimislaud on ärianalüüsi vahendi väljund, mis kuvab ühe konkreetse valdkonna kriitilised mõõdikud või näitajad eraldiseisva vaatenähtena, võimaldades kasutajal jälgida ning uurida ettevõtte eesmärkide täitmise kulgu.

Masinõpe (*Machine Learning*) – programmid, millel on võime õppida „kogemustest“ (suuremast andmehulgast) ja väljastada paremaid tulemeid.

Metaandmed (*Metadata*) – andmed andmete kohta ehk andmed, mis kirjeldavad andmete erinevaid tahke ja omadusi. Metaandmeid kasutatakse ärianalüüsi lahendustes andmete

pärimisel (tehnilised metaandmed) ning andmete sisu üheseks ning täpselt mõistmiseks (ärilised metaandmed).

ODS (*Operational Data Store*) – andmelaos asuv lähtesüsteemi andmete koopia.

Relatsiooniline andmebaas (*Relational Database*) – tabelite, kirjete ja veergude kaupa korraldatud andmete kogum [1]. Relatsiooniline andmebaas loob täpselt määratletud seosed andmebaasi eri tabelite vahel. Tabelite vahel määratud seosed hõlbustavad andmete otsimist, korraldamist ja aruandlust. Relatsioonilises andmebaasis kasutatakse struktureeritud päringukeelt (SQL). Relatsioonilise andmebaasi juhtimissüsteemideks on näiteks Oracle, MySQL, PostgreSQL, SAP Sybase.

Struktureeritud andmed (*Structured Data*) – andmed, mille struktuur (ülesehitus) ning koosseis on enne andmete kasutuselevõttu kokku lepitud. Struktureeritud andmeid on võimalik salvestada relatsioonilisse andmebaasi ning sealt pärida struktureeritud päringukeele abil.

Struktureerimata andmed (*Unstructured Data*) – andmed, mille struktuur (ülesehitus) ning koosseis ei ole enne andmete kasutuselevõttu kokku lepitud. Struktureerimata andmed on üldjuhul muutuva struktuuri ja koosseisuga.

STG (*Staging*) – andmete cellaadimise ala. See on kiht, kuhu regulaarselt imporditakse lähtesüsteemidest uusi andmeid, ehk pärast viimast andmelao laadimist lähtesüsteemi lisatud ja/või muudetud andmeid.

SQL (*Structured Query Language*) – päringukeel relatsiooniliste andmebaasidega suhtluseks.

NoSQL (*Not only SQL*) andmebaas – on vastand relatsioonilisele SQL andmebaasile. NoSQL andmebaas lähtub mitterelatsioonilise andmebaasi põhimõttest ning erinevalt relatsioonilisest andmebaasist ei pea NoSQL süsteemid vastama rangelt ette kirjutatud andmemudelitele. Sellest tulenevalt suudavad NoSQL andmebaasid käsitleda mistahes andmeid [1].

1. Analüütika valdkonna tutvustus

Käesolevas peatükis tehakse sissejuhatus analüütikavaldkonda ning selgitatakse peamisi magistritöös käsitletavaid mõisteid. Kesksel kohal on ärianalüüs kui analüütikaprojektide konteksti läbivaim mõiste. Peatükis selgitatakse ärianalüüsi olemust, eesmärki ning ärianalüüsi rakendamise võimalusi. Sissejuhatusena analüütikaprojektide realiseerimise tehnoloogilistesse lahendustesse tutvustatakse tehnilisi lahendusi, mille abil saab ärianalüüsi projekte ellu viia. Peatüki lõpetab ärianalüüsi trendide ülevaade sellest, mis suunas analüütikalahendused on liikumas.

1.1 Ärianalüüsi ning ärianalüütika mõiste

Otsuste, järelduste ning statistilise analüüsi tegemiseks on vaja korrektseid andmeid ning oskust neid õigesti tõlgendada. Analüütikaprojektide kontekstis on läbivaks mõisteks ärianalüüs¹ (ingl *Business Intelligence* – *BI*). Ärianalüüs on tehnoloogiliste vahendite kasutamine ettevõtte andmete defineerimiseks, uurimiseks ja analüüsimiseks [1]. Näitena saab siin tuua müügitulu, toodete või kulude analüüsimise tehnoloogiliste lahenduste abil.

Teine oluline mõiste analüütika valdkonnas on ärianalüütika (ingl *Business Analytics* – *BA*). Kui ärianalüüsi eesmärgiks on teha andmetele tuginedes paremaid juhtimisotsuseid, siis ärianalüütika eesmärgiks on anda ärianalüüsiks vajalikud andmed ning meetodid. Seega on ärianalüütika kogum lahendustest, mida rakendatakse analüüsimodelite kasutamiseks, mõistmaks tegelikkust ja prognoosida tulevikku [2]. Selle mõiste alla kuuluvad näiteks andmekaeve, ennustav analüüs ja statistiliste mudelite ehitamine.

Ärianalüüsiks vajab ettevõtte ennekõike andmeid ning vastavat tarkvara, mis koondab kokku erinevates allikates sisalduvad andmed ja võimaldab juhtidele luua tähendusrikkaid visuaale ning juhtimislaudu (ingl *Dashboard*).

1.2 Ärianalüüsi eesmärk

Ärianalüüsi eesmärgiks on anda ettevõttele vahendid, mis aitavad ettevõttel areneda, olla konkurentsivõimelisem ning teha kaalutletud otsuseid. Seega võimaldab ärianalüüs ettevõttele tõhusama otsustusprotsessi ja strateegilise tegevuse.

Ärianalüüs kasvab ettevõtet toetavate lahendustega pidevalt ning kohandub paralleelselt tehnoloogia arenguga. Kaasaegse ärianalüüsilahenduse osadeks on [3]:

¹ Täiendavalt on ärianalüüsi mõiste Eestis kasutusel äriprotsesside kaardistamisel.

- reaalaajas andmelaondus,
- andmekaeve,
- automaatsed anomaaliate ja erindite tuvastused,
- proaktiivsed teavitused koos teavituse saaja automaatse tuvastusega,
- sujuva töövoogu tagamine,
- masinõppe kasutamine,
- geoinfosüsteemid,
- andmete visualiseerimine.

Ärianalüüs kasutab ettevõtte ärivajadusest lähtuvalt erinevaid tehnoloogilisi lahendusi, et tagada ettevõttele juurdepääs andmetele, mis on kriitilised allikad õigete otsuste tegemiseks. Lisaks konkurentsivõime kasvamisele võimaldab ärianalüüs vähendada ettevõtte kulusid, tõsta tulusid ning klientide rahulolu, parandada ettevõtte andmete kvaliteeti ja palju muud.

1.3 Ärianalüüsi võimalused

Ärianalüüsi rakendamine annab ettevõtetele erinevaid võimalusi alustades lihtsatest aruannetest ja juhtimislaudadest ning lõpetades masinõppel baseeruvate prognoosimudelitega. Sarnaselt erinevatele analüüsimeetoditele võib ka ärianalüüsil olla erinevaid tarbijaid.

Tarbijateks võivad olla ettevõtte juhid (teevad äriotsuseid) või spetsialistid (vajavad ülevaateid valdkonna seisust) jne. Lisaks ei pea ärianalüüsi tulemi tarbijaks olema alati inimene. Analüütikalahendus võib olla sisendiks ka infosüsteemile. Näitena saab siin tuua:

- täiendava info kuvamine operatiivlahendustes (näiteks analüütikalahenduses arvatud kliendi riskiskoor panga telleri töövahendis);
- automatiseeritud tellimuste tegemine vastavalt analüütikalahenduse poolt prognoositavale laovarude lõppemisele ja tarnetähtaegadele.

Ärianalüüsi võimalused on väga laiad. Leidub erinevaid paindlikke tehnoloogiaid, mis võimaldavad andmeid ettevõtte vajadustest lähtuvalt enda kasuks tööle panna. Seejuures võib ärianalüüsi kasusaajaks olla nii inimene kui ka tehniline tarbija (näiteks infosüsteem).

1.4 Ärianalüüsis kasutatavate andmete tüübid

Andmed, mida analüütikaprojektides kasutakse, on kas struktureeritud või struktureerimata. Need on mõisted, mida ärianalüütika kontekstis läbivalt kasutatakse ning mis leiavad kajastust ka käesolevas magistritöös.

Struktureeritud andmete korral on andmete koosseis ja struktuur eelnevalt kokku lepitud. Neid on võimalik salvestada relatsioonilisse andmebaasi ning pärida sealt struktureeritud päringukeele (ingl *Structured Query Language – SQL*) abil [1].

Struktureerimata andmed on pidevalt muutuva struktuuri ja koosseisuga. Struktureerimata andmeid ei ole mõistlik salvestada relatsioonilisse andmebaasi, vaid selleks on olemas spetsiaalsed NoSQL (*Not only SQL*) andmebaasid. Tüüpiliselt käsitletakse struktureerimata andmetena näiteks dokumendifaile (MS Word, PDF jt dokumendid), audiofaile, sõnumivahetuse andmeid jpt [1].

1.5 Ärianalüüsi projektide tehnilised lahendused

Analüütikavaldkonna enamlevinud tehnilisteks lahendusteks on andmeladu, andmejärv ja andmejaotur² [4]. Eestis kasutatakse neist enim andmeladu. Viiteid andmejärve ning andmejaoturi Eestis kasutamise kohta magistritöö autor ei ole leidnud.

Kõik eelmises lõigus nimetatud tehnoloogiad koosnevad mitmest komponendist. Suur osa komponentidest on lõppkasutajate eest peidus ning nende eesmärgiks on tagada nõuetele vastavate andmete kättesaadavus välisele tarbijale. Peamiseks väliseks tarbijaks on lihtsa kasutajaliidesega juhtimislaud või aruandlus- ja analüütikakeskkond. Üldistatult võib neid nimetada andmete presenteerimise komponendiks. Eestis kõige tuntumad andmete presenteerimise töövahendid on Power BI [5] ja Tableau [6].

1.6 Ärianalüüsi trendid

Vahendid, mida ärianalüütikaks kasutatakse, peaksid toetama valdkonna trende. Järjest rohkem liiguvad lahendused andmete reaajas laadimise suunas. Ettevõtetele ei piisa enam kord kuus või isegi kord päevas tehtavatest analüüsides, vaid erinevatest süsteemidest peab olema võimalik andmeid koondada ja analüüsida oluliselt sagedamini.

Lisaks on andmete ja analüütika trendid suundumas lahendustele, mis võimaldavad [7]:

- lisaks struktureeritud andmete kasutamisele analüüsida ka struktureerimata andmeid;

² Autori eestindus.

- pilvepõhiste lahenduste kasutamist, tagamaks paindlikumat laiendatavust ja jõudluse kasvu;
- graafiliste väljundite laiemat kasutamist, sh tehisintellekti ning masinõppe (ingl *Machine Learning*) kasutamist selgemate ja paremini tõlgendatavate graafikute loomiseks;
- dünaamiliste, konkreetse kasutaja vajadustest ja soovidest lähtuvate juhtimislaudade loomist.

Sõltumatu ning objektiivne uuringufirma Gartner [8], on 2021. aasta andmete ja analüütika trendidest toonud välja järgmise arutelu [7]: Üha rohkem ettevõtte juhte hakkavad mõistma andmete ning analüütika tähtsust ettevõtte kasvu suurendamise kontekstis. Kui senini on andmeteadlased/analüütikud olnud ettevõttes eraldiseisev meeskond ning teisejärgulisel positsioonil, siis tulevikus liigub andmete analüüsimise roll ettevõtte võtmetegevuste alla. Siiski alahindavad ettevõtte juhid andmete ning nende tõlgendamise ja analüüsimise keerukust. Kui andmehalduse juhte (ingl *Chief Data Officer – CDO*) kaasata ettevõtte eesmärkide ja strateegiate otsustusprotsessi, suudaksid nad järjepidevalt suurendada ettevõtte tootlikkust ja äriväärtust kuni 2,6 korda.

Konkreetsete tehnoloogiliste lahenduste trende vaadeldes võib näha ettevõtete selgeid eelistusi. Gartneri 2020. aasta uuringu kohaselt investeerivad IT- ettevõtted andmete ning analüütikalahenduste valdkonnas kõige enam just andmelao, andmejärve ning andmejaoturi tehnoloogiasse [4].

Uringust nähtub, et enamus ettevõtetes on kasutusel andmeladu, kuid prognooside kohaselt on järgneva kahe aasta lõppedes kõik kolm (andmeladu, andmejärv ning andmejaotur) tehnoloogiat kasutusel suhteliselt võrdsel määral.

Analüütikavaldkonda tutvustavas peatükis selgitati ärianalüüsi eesmärki ning rolli, mida ärianalüüs täidab ettevõtte edukuses. Ärianalüüs muutub ettevõtte juhtidele üha olulisemaks ning võimalusi ärianalüüsi rakendamiseks on rikkalikult.

Järgnevas peatükis keskendutakse ärianalüüsi realiseerimise võimalustele. Võimaluste uurimisel analüüsitakse ja kirjeldatakse analüütikaprojektide realiseerimise tehnoloogiad.

2. Analüütikaprojektide realiseerimise tehnoloogilised lahendused

Käesoleva peatüki eesmärgiks on analüüsida ja kirjeldada, millised on peamised analüütikaprojektide tehnoloogilised lahendused, mida kasutatakse suuremate projektide korral. Peatükis käsitletakse andmeladu, andmejärve, loogilist andmeladu ning andmejaoturit. Välja tuuakse peamised mõisted, arhitektuur, eelised ning puudused.

Analüütikaprojektide realiseerimise tehnoloogiaid analüüsitakse ja kirjeldatakse käesolevas peatükis suurte lahenduste vaatest. Suur lahendus on selline, mis kasutab mitmeid erinevaid andmeallikaid, erinevaid andmeformaate, erinevaid andmetele ligipääsemise meetodeid ning lahendusel on mitmeid erineva profiiliga kasutajaid. See tähendab, et suurel analüütikalahendusel on väga palju erinevaid nõudeid/vajadusi. Samas leidub nende nõuete täitmiseks mitmeid erinevaid tehnoloogiaid.

Tehnoloogiate analüüsil lähtutakse selles peatükis järgmistest küsimustest:

- millised on lahenduses kasutatavad andmed;
- kes ja milleks neid andmeid kasutab;
- kuidas andmeid talletatakse;
- kuidas liigutatakse andmeid analüütikalahendusse ja selle sees.

Nendele küsimustele saab koondatud vastused iga tehnoloogilise lahenduse alapeatüki kokkuvõttes lõigus.

2.1 Andmeladu

Andmeladu (Eestis on sünonüümina kasutusel ka mõiste andmeait) koos andmeid visualiseeriva kasutajaliidesega on ettevõtte juhtidele ning spetsialistidele peamine töövahend, mis tagab juhtimisotsusteks vajaliku info.

Lihtsustatult võib andmeladu kirjeldada kui infosüsteemi, mis koondab andmed erinevatest keskkondadest ning võimaldab koostada aruandeid ning teostada nendel andmetel tuginevat analüüsi.

Ärianalüüsi projektide tehniliste lahenduste jaotis 1.5 tõi välja, et ärianalüüsi lahenduse tehnoloogiad (sh andmeladu) koosnevad mitmest komponendist ning suur osa komponentidest on lõppkasutajate eest peidus.

Vaatleme andmeladu ning selle komponente lähemalt. Andmelao lahendus tervikuna sisaldab [2]:

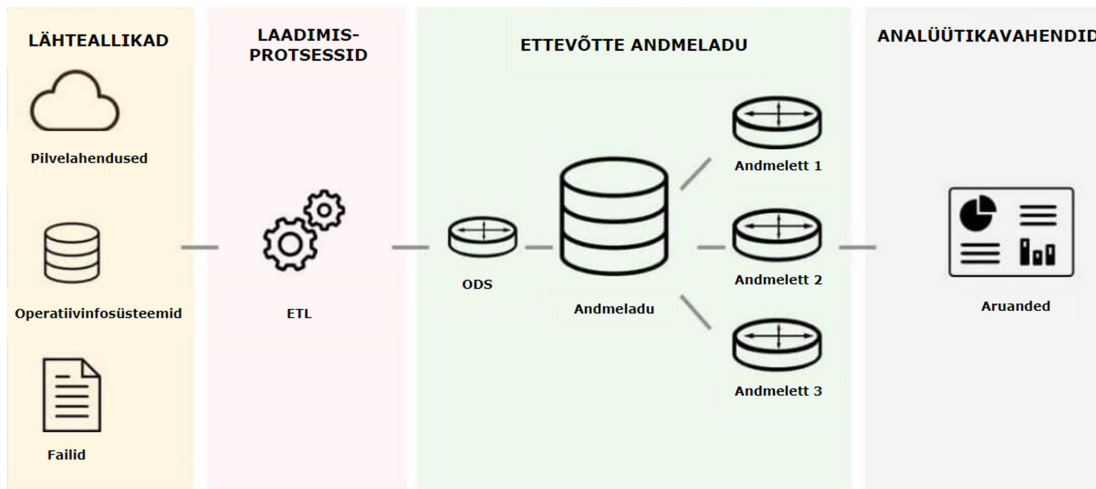
- andmete lähtesüsteeme (operatiivinfosüsteeme);
- andmete laadimise protsessi (reeglina koos andmete ühtlustamise ja puhastamisega);
- andmebaasi/andmeid (summeeritud andmed, detailandmed);
- aruandlusvahendeid (andmete visualiseerimine, analüütilahendused);
- andmelao halduskeskkonda, sh metaandmeid (tehnilised metaandmed, ärilised metaandmed).

Ülaltoodud loetelust on andmete lähtesüsteemid andmelao lahenduse andmeallikateks. Lähtesüsteemideks võivad olla näiteks tootmistarkvara ja kliendihaldustarkvara. Ülaltoodud loetelust kolm komponenti: andmete laadimise protsess, andmebaas ning halduskeskkond, on aga andmelao arendusmeeskonna ehitada ja hallata. Need kolm komponenti on andmelao lõppkasutaja eest nõ peidus. Andmelao komponentidest on lõppkasutajale nähtav ehk kasutajaliidese kaudu kättesaadav aruandlusvahendi komponent.

Andmelao mõistet on võimalik käsitleda nii laiemas kui ka kitsamas vaates. Andmeladu kitsamas vaates on andmebaas, mis on optimeeritud aruandlusele ja analüüsi teostamisele [9]. Andmeladu laiemas vaates sisaldab andmelao mõiste erinevaid protsesse, meetodeid ja tehnoloogiaid andmete kasutamiseks, sealhulgas ka andmelao andmebaasi, lähtesüsteeme ja aruandlusvahendeid.

Seda, kas andmeladu on käsitletud kitsas või laias mõttes, defineerib ära kontekst, milles andmeladu on esitletud. Käesolevas magistritöös lähtutakse peamiselt andmelao laiemast mõistest.

Andmelao arhitektuuri kirjeldavatel joonistel on andmeladu tihti seostatud andmebaasi tähistava ikooniga. Selgitava näitena on lisatud allolev joonis 1 kui üks tavapäraseid lähenemisi, kuidas esitatakse andmelao arhitektuuri kirjalikes allikates [10]:



Joonis 1. Andmelao arhitektuuri esitamise näidis

Joonisel 1 on kujutatud andmelao arhitektuur. Rohelisel alal on kujutatud andmeladu kitsamas mõttes ehk andmeladu kui andmebaas. Koos teiste joonisel toodud komponentidega on tegemist andmelaoga laiemas mõttes.

Andmelaol on neli tähtsat omadust, millest andmeladu ehitades alati lähtutakse. Omadused on defineerinud W.H. Inmon, kelle The Global Data Management Community on tituleerinud andmelao isaks [11]. Inmon kirjeldab andmelao olulisi omadusi ning defineerib, et andmeladu on [12]:

- subjektorienteeritud (ingl *Subject Oriented*),
- integreeritud (ingl *Integrated*),
- püsiv (ingl *Nonvolatile*) ning
- ajast sõltuv (ingl *Time Variant*) andmebaas.

Subjektorienteeritud – st andmeladu on orienteeritud ettevõtte tegevusvaldkondadele, sündmustele (nt operatiivinfosüsteemid on objektorienteeritud ehk keskenduvad objektidele nagu klient, teenus, arve jt).

Integreeritud – kogutakse kokku lähtesüsteemidest saadavad andmed, mis ühtlustatakse, puhastatakse ning korrastatakse.

Püsiv – andmelattu laaditud algandmed hoiustatakse nende algupärasel kujul. Ajast sõltuv – võtmestruktuuriks on aeg. Andmete analüüsi tehakse ajast (aasta, kuu jt) lähtuvalt st andmed on otseses seoses ajaga.

Eestis kasutusel olevatest andmelao lahendustest võib näitena tuua Eesti tervise infosüsteemi (TIS) andmeid analüüsiva statistika andmelao [13] ning töötukassa andmelao (TARU) koos selle peale ehitatud otsusetoega OTT. OTT prognoosib töötu tööle liikumise

töenäosust, toob välja seda mõjutavad tegurid ja aitab seeläbi pakkuda töötule abi individuaalsetest vajadustest lähtuvalt ning tõsta töötukassa efektiivsust [14].

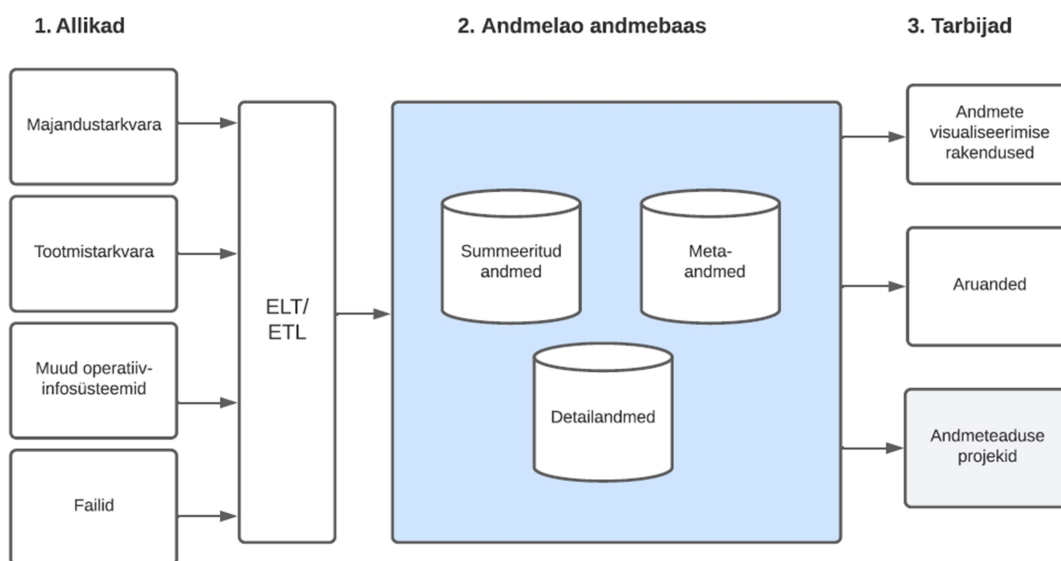
2.1.1 Andmelao arhitektuur

Käesolevas jaotises selgitatakse andmelao tööpõhimõtet. Sellest annab kõige parema ettekujutuse allpool esitatud arhitektuuriline joonis, kus on välja toodud andmelao olulised komponendid. Joonise 2 alla on lisatud komponentide selgitused ning komponentide rollid.

Andmed on andmelao peamine "kütus". Andmed voolavad andmelattu erinevatest infosüsteemidest/andmebaasidest (näiteks majandustarkvarad, kliendihaldusprogrammid jpt) ja välistest allikatest nagu partnerite süsteemid. Andmed kontrollitakse, puhastatakse, teisendatakse ja seostatakse omavahel, et muuta nad omavahel võrreldavaks ja töödeldavaks. Seejärel andmed salvestatakse andmelattu, mille järel on andmed kasutusvalmis.

Andmelao arhitektuuri selgitamisel on aluseks võetud Resta OÜ (edaspidi Resta) andmelao arhitektuur. Resta on enam kui 30 aastat tegutsenud andmeladude ja ärianalüütikalahenduste arendaja [15] ning omab praktilist kogemust kümnete suurte projektide näol, seeläbi sobitub ettevõtte valik magistritöö eesmärgiga: analüüsida analüütikalahendusi suurte projektide perspektiivist vaadatuna.

Resta ehitatud andmeladude näitel on andmelao arhitektuur järgmine [9]:



Joonis 2. Andmelao arhitektuur andmete liikumise põhimõttest lähtudes

Arhitektuurijoonisest nähtub, et moodustub kolm peamist gruppi: allikad, andmelao andmebaas ning tarbijad. Allikatest liiguvad andmed andmelao andmebaasi, kust tarbijad saavad aruannete või visualiseerivate rakenduste jaoks vajalikke andmeid.

Vaatleme kolme gruppi ning neid esindavaid komponente lähemalt:

1. Andmeallikateks võivad olla erinevad ettevõttesisesed ja -välised infosüsteemid/ andmebaasid (näiteks majandustarkvarad, kliendihaldusprogrammid jpt), veebilehed, sotsiaalmeedia ja partnerite süsteemid. Samuti võivad andmeallikateks olla failid ja dokumendid nagu näiteks XML-failid, MS Exceli tabelid jms.

2. ELT või ETL³ lahendusega koondatakse andmed lähteallikatest andmelao andmebaasi.

ELT ja ETL on mõlemad protsessid, mis vastutavad andmete laadimise, töötlemise ning andmebaasi sisestamise eest. Sisuliselt on ETL ja ELT näol tegemist sama lahendusega, erineb vaid vastavate alamprotsesside tööjärjekord. ELT ja ETL on saadud ingliskeelsete sõnade *Extract, Transform ja Load* esitähedest.

Eraldamise (E) faasis imporditakse andmed lähteallikast andmete laadimise protsessiga tegelevasse süsteemi. Lisaks valitakse selles etapis välja, mis andmed vastavast süsteemist on vaja andmelattu transportida.

Teisendamise (T) faasis teisendatakse andmed aruandluseks ja analüüsiks sobivale kujule. Lisaks toimub andmete ühtlustamine, mille käigus võetakse kasutusele ühesugused väärtused üle kõigi lähtesüsteemide. Sealhulgas toimub andmete puhastamine, mille käigus eemaldatakse näiteks vigased andmed ja duplikaadid.

Laadimise (L) faasis laaditakse eelmises faasis teisendatud andmed andmelao andmebaasi. Andmelattu salvestatavad andmed võib jagada detailandmeteks (näiteks üksikud müügitehingud) ja summeeritud (agregeeritud) andmeteks (näiteks kuu müügikoondid). Summeeritud andmeid kasutatakse enamlevinud päringute kiirendamiseks. Lisaks sisulistele andmetele salvestatakse andmelao andmebaasi ka metaandmed ehk andmed andmete kohta. Kõiki neid erinevaid andmetüüpe kasutatakse tarbijatele vajalike analüütikatulemite loomiseks.

Metaandmed kirjeldavad andmelao andmeid nii tehniliselt (millised tabelid ja väljad on andmebaasis, millised on andmetüübid jne) kui ka äriselt (mis on vastava andmevälja

³ ETLi loetakse tänapäeval juba vananevaks tehnoloogiaks ning enamasti on kasutusel ELT loogika.

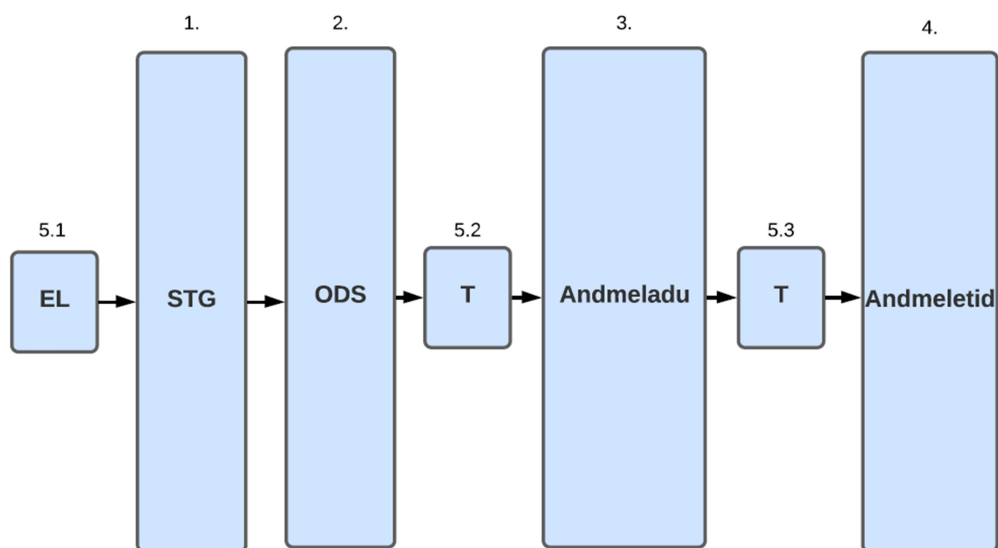
äriline tähendus). Enamik ärianalüüsi tarkvarasid kasutavad andmelao metaandmeid selleks, et juhtimislaudade ja/või aruannete koostajad ei peaks teadma andmelao täpset tehnilist andmemudelit (tabelid, väljad, seosed), vaid saaks oma töövahendis kasutada vastavaid ärilisi termineid.

3. Tarbijateks on andmelao puhul peamiselt ärikasutajad. Ärikasutajatele teeb andmelaos ettevalmistatud andmed kättesaadavaks andmete visualiseerimise rakendus (näiteks Tableau, MS Power BI), juhtimislaud või mõni muu analüütikarakendus. Võimalikeks andmelao tarbijateks on ka andmeteaduse projektid. Joonisel on need markeeritud halli taustaga, kuna andmeteaduse projekt esineb andmelao tarbijana oluliselt vähem kui aruanne või andmete visualiseerimise rakendus.

Järgmisena liigume andmelao arhitektuuris ühe taseme võrra sügavamale. Vaatleme, milline on andmelao andmebaasi sisemine ülesehitus (joonisel 2 sinise värviga tähistatud 2. Andmelao andmebaas) ning milliseid tegevusi andmetega teostatakse.

Resta on andmelao andmebaasi ülesehitamisel lähtunud suurtest projektidest, mis esitavad andmelao arhitektuurile kõrged nõudmised. Vastavalt Resta metoodikatele jaotatakse andmelao andmebaas mitmeks kontseptuaalseks kihiks. Iga andmebaasi kiht omab kindlat rolli vajalike analüütikatulemite loomisel. Kihiline andmebaasi arhitektuur võimaldab tagada lahendusele esitatud kõrged ärinõuded nagu veakindlus, kiire aruandlus, lahenduse laiendatavus, optimaalne lähtesüsteemide ressursikasutus jpt.

Andmelao andmebaasi arhitektuuri kihid on kujutatud joonisel 3.



Joonis 3. Andmete liikumine andmelao andmebaasis ELT laadimismeetodi näitel

Joonise 3 selgitused (komponendid 1-5) [9]:

1. STG (ingl *Staging*) – andmete eellaadimise ala. See on kiht, kuhu lähtesüsteemidest imporditakse regulaarselt uusi andmeid (peamiselt pärast viimast andmelao laadimist lähtesüsteemi lisatud ja/või muudetud andmed).

Roll: STG kihi roll on ajutiselt hoiustada lähtesüsteemist imporditud andmeid ning kontrollida nende terviklikkust enne nende ülekandmist järgmisse kihti. Andmed laaditakse lähtesüsteemidest STG alasse. Laaditakse vaid kokkulepitud andmed, kokkulepitud andmestruktuuriga ning kokkulepitud ajal. Laadimisel kontrollitakse, kas andmed vastavad kokkuleppele ning kas neid on võimalik kasutada. Kui laadimine ja kontrollid õnnestuvad, on andmed valmis transpordiks järgmisse andmelao kihti – ODS-i. STG kihis hoiustatud andmed kustutatakse pärast andmete edukat ülekandmist ODS kihti.

2. ODS (ingl *Operational Data Store*) on sisuliselt koopia lähtesüsteemist. ODS-is on kasutusel täpselt sama andmemudel mis lähtesüsteemis (samad tabelid, andmeväljad, seosed jne).

Roll: Kui STG kihis hoiustatakse vaid uusi andmeid, siis ODS kihis säilitatakse kõik andmelattu laaditud andmed (kogu andmete ajalugu). Nii STG kui ODS kihi andmemudel on sama, mis on lähteallika andmemudel. ODS kiht võimaldab hõlpsalt võrrelda lähteallika ja andmelao andmebaasi andmeid ning seda saab kasutada laadimisvigade tuvastamiseks. Samuti on ODS kihti võimalik kasutada lihtsamate aruannete genereerimiseks.

3. Andmeladu – äri vajadustest tuleneva (aruandluse ja analüüsi jaoks optimeeritud) andmemudeliga andmete hoiustamise ja töötlemise kiht. Andmelaos olevad andmed on sellisel kujul ning sellise detailsusega, mis võimaldavad hõlpsalt teha ärianalüüsi.

Roll: Andmelao ülesanne on organiseerida lähteallikast imporditud andmed ümber ning viia need analüüsiks ja aruandluseks sobivale kujule. Andmed laaditakse ODS kihist andmelattu läbi teisendamise faasi (joonisel 5.2). Andmelao laadimise käigus ODS-ist saadavad andmed puhastatakse, standardiseeritakse, rikastatakse (st lisatakse täiendavaid andmeid teistest lähtesüsteemidest) ning tehakse muu vajalik eeltöö aruannete tegemiseks. Andmelao kihti kasutavad ärianalüütikud andmete analüüsimiseks (ärianalüüsi teostamiseks).

4. Andmelett on ühe konkreetse äriülesande lahendamiseks mõeldud andmestik (võib koosneda ka mitmest andmetabelist).

Roll: Andmelette kasutatakse eelkõige sagedasti kasutatavate (standard)aruannete alusandmetena. Tihti kuuluvad siia summeeritud andmetega tabelid (agregeeritud tabelid). Mitmeid aruandeid saab genereerida otse andmelao andmemudelilt ning selles sisalduvate andmetega ilma, et oleks vaja andmeletti. Andmeletti on vaja ennekõike juhtudel, kus ettevõtte vajab aruandesse summeeritud, eelnevalt väljaarvutatud andmeid. Üks näide, kus andmelett võib otstarbekas olla, on aruanne, mis näitab kuu lõpu seisuga kõigi ettevõttes tehtud tehingute summat. Selliste arvutuste tegemine otse andmelao andmemudeli pealt oleks andmelaole liigselt koormav ning võib tekitada aeglust nii olemasoleva kui ka teiste aruannete töös.

5. ELT (ingl *Extract, Load, Transform*) – andmete eraldamise, laadimise ning teisendamise protsess.

Vastavalt Resta metoodikale kasutatakse ELT laadimismeetodit. Andmete teisendamine ei toimu ühe protsessina, vaid on jagatud mitmeks osaks. Joonisel 3 nähtub, et siin kasutatakse $EL+T=ELT$ loogikat. See tähendab, et andmete teisendamine teostatakse andmelao andmebaasi sees (tüüpiliselt SQL lausetena või andmebaasi protseduuridena).

Roll: ELT on protsess, mille ülesandeks on andmete liigutamine ühest kihist teise ning vastavalt vajadusele ka andmete teisendamine.

Joonisel 3 kirjeldatud numeratsioonile tuginedes on ELT protsess järgmine:

5.1. Esmalt eraldatakse lähtesüsteemidest andmelao jaoks vajalikud andmed ning laaditakse kõigepealt STG-kihti ning sealt edasi ODS kihti. Kuna andmete transportimisel STG kihist ODS kihti sisulist andmete teisendamist ei toimu, siis ei ole seda joonisel eraldi teisendusprotsessina välja toodud.

5.2. Seejärel toimub andmete teisendamine aruandluseks sobivale kujule.

5.3. Viimase sammuna toimub andmete teisendamine andmelaost andmelettidesse.

ELT kirjeldamine oli joonisel 3 esitatud andmelao andmebaasi kihtidest viimane, ning seega on kõik andmelao arhitektuuri komponendid ja rollid selgitatud. Ülaltoodud andmelao arhitektuur oli esitletud suure ning keeruka projekti näitel. Lihtsama projekti korral ei ole kõiki siin jaotises tutvustatud kihte vaja realiseerida ning mitmeid erinevaid tegevusi võib teostada ühe protsessi sees. Seega on iga andmelao arhitektuur erinev ning kujuneb konkreetse ettevõtte ärivajadustest.

2.1.2 Andmelao tugevused

Eelnevalt vaadeldi andmelao üldist olemust ning arhitektuurilist ehitust. Andmelaost terviklikuma ülevaate saamiseks käsitletakse käesolevas jaotises andmelao kui analüütikalahenduse tehnoloogia tugevusi.

Andmelaol on järgmised tugevad omadused [16]:

- võimaldab ettevõttel teha strateegilisi otsuseid ning juhtida efektiivsemalt, andes sellega kaudse eelise konkurentide ees;
- võimaldab ligipääsu ajaloolistele⁴ andmetele;
- tagab standardiseeritud ja kvaliteetsed andmed;
- suurendab andmeanalüüsi võimekust ja kiirust;
- tõstab ettevõtte müügitulu ja tugevdab ettevõtet läbi juhtimisotsuste, mis põhinevad ajakohastel ning korrektsetel andmetel;
- kasutamisevõimalused on hõlpsalt laiendatavad ettevõtte vajadustest lähtuvalt;
- on kasutatav nii ettevõtte lokaalses infrastruktuuris kui ka pilvepõhise teenusena.

Võrreldes teiste analüütikalahenduste tehnoloogiatega, tõuseb andmeladu esile ennekõike kvaliteetsete, kontrollitud andmete vahendajana.

2.1.3 Andmelao nõrkused

Vastukaaluks eelmises jaotises vaadeldud andmelao tugevustele, on loetletud andmelao peamised nõrkused [17]:

- Andmelaol on kõrged hoolduskulud (nii infrastruktuuri, kvalifitseeritud meeskonna kui ka võimalike tarkvara litsentside kulud).
- Kuigi andmelad töötavad suurepäraselt struktureeritud andmete korral, on struktureerimata andmetega töötamisel eelised teistel tehnoloogilistel lahendustel.
- Olukorras, kus ettevõtte ärivajadused kiiresti muutuvad, võib spetsiifilise äriprobleemi lahendamise jaoks koostatud andmelao aruandlus muutuda aegunuks.
- Andmelao laadimise protsess (andmete eraldamine lähtesüsteemides, teisendamine ja laadimine andmelao andmebaasi) võtab paratamatult aega ja seetõttu ei kajasta andmelao aruandlus üldjuhtudel käesolevat hetkeseisu.

Andmelao (nagu ka teiste analüütikalahenduste) loomine on suur arendusprojekt, mille juurutus vajab nii arendusmeeskonna kui ka ettevõtte spetsialistide ning juhtide olulist

⁴ Lähtesüsteemidest võidakse süsteemi jõudluse säilitamiseks vanemad andmed ära kustutada (näiteks vanemad kui 2a vanused andmed). Andmelattu imporditud andmeid ei kustutata.

panust. Efektiivse andmelao loomise võtmeteguriks on projekti tellija ja täitjaga koostöös sündinud ettevõtte äri vajadusi arvestav andmelao andmemudel.

2.1.4 Andmelao kokkuvõte

Andmelao peatükis selgitati andmelao olemust ning definitsiooni. Andmeladu käsitletakse ennekõike andmelao mõiste laiemas määratluses – andmeladu kui tehnoloogiate ja praktikate kompleks, mis tagab juhtimisotsusteks vajaliku info. Erinevate jooniste toel toodi välja andmelao arhitektuur, komponendid ning komponentide rollid. Lisaks kirjeldati peamisi andmelao eeliseid ja puuduseid.

Andmete perspektiivist vaadatuna, saab andmelaost tehtud ülevaate kokku võtta järgmiselt:

- andmeladu kasutab peamiselt struktureeritud andmeid;
- andmelao peamiseks kasutajateks on ärikasutajad, erinevad spetsialistid ning ettevõtte juhid;
- andmed on salvestatud andmelao andmebaasi ning
- andmed liiguvad andmelaos läbi ETL või ELT lahenduse.

Andmeladu on käesolevas magistritöös kirjeldatud analüütikaprojektide tehnoloogiatest kõige vanem ning ettevõtetes kasutusel juba enne sajandivahetust. Järgnevates peatükkides käsitletud tehnoloogiad on oluliselt uuemad. Mõned uute tehnoloogiatega kaasnevad mõisted ei ole veel eestikeelses terminoloogiasse jõudnudki.

2.2 Andmejärv

Eelmises peatükis käsitletud andmeladu on eelkõige mõeldud struktureeritud andmete töötlemiseks ja analüüsiks. Samas tänu asjade internetile (ingl *Internet of Things* – IoT) ja sotsiaalmeedia plahvatuslikule arengule tekib järjest rohkem struktureerimata andmeid. Andmeid oskuslikult kasutades on võimalik saada palju väärtuslikku infot. Struktureerimata andmete kasv aga tähendab, et andmeladu ei pruugi analüütikalahenduse loomiseks olla tänapäeval piisav tehnoloogia.

Üks võimalikke lahendusi on andmejärv (ingl *Data Lake*). Andmejärv on tehnoloogiline lahendus mis võimaldab koguda, säilitada, töödelda ja tagada kasutajatele ligipääsu mistahes (st nii struktureeritud kui ka struktureerimata) andmetele.

Andmejärve mõistet mainis esimest korda 2010. aastal James Dixon, kes kõrvutas andmejärve mõistet andmelettidega [18]. Dixon nägi andmejärve kui suurt kogumit, kuhu

andmed voolavad sisse eri allikatest, ning kust erinevad kasutajad saavad andmeid analüüsida või kasutusele võtta.

Andmejärve mõiste on uudne ning kirjalikke materjale napib. Autori poolt leitud kirjalikes allikates on andmejärve temaatikat käsitletud peamiselt mõiste ning olemuse tasemel. Kõige värskemalt ning põhjalikumalt on andmejärve analüüsinud 2020. aastal Sawadogo ning Darmont, kes on andnud andmejärvele järgneva definitsiooni [19:100]:

„Andmejärv on hõlpsalt laiendatav andmete hoiustamise ja analüüsi süsteem, mis hoiustab mistahes tüüpi andmeid. Andmejärves säilitatakse andmed nende algses vormingus ning kasutatakse peamiselt andmespetsialistide (statistikud, andmeteadlased, analüütikud) poolt teadmiste ammutamiseks“.

Andmejärve tehnoloogia sisaldab [19]:

- mistahes tüüpi andmete kogumist,
- andmete organiseerimist ja säilitamist,
- andmehaldusvahendeid ning -poliitikat,
- metaandmeid,
- juurdepääsu erinevatele kasutajatele,
- laiendatavust andmete hoiustamiseks ja töötlemiseks.

Erinevalt andmelaost, ei tegele andmejärve tehnoloogia konkreetsete äriprobleemide lahendamise jaoks vajalike andmete ettevalmistamisega. Andmejärve lahenduse fookus on erinevatest allikatest mistahes tüüpi andmete kokku kogumisel ning andmete säilitamisel nende algupärasel kujul. Enamik andmejärve lahendusi ei sisalda andmete teisendamist või siis tehakse teisendusi vaid vähesel määral.

Eestis on andmejärve tehnoloogia uudne. Kui andmelao lahendusena realiseeritud projektidest võis eelmises peatükis välja tuua mitmeid näiteid, siis Eestis realiseeritud andmejärve lahenduse kohta selgeid kirjalikke allikaid töö autor ei ole leidnud.

Mujal maailmas on andmejärv laiemalt kasutusel mitmetes riikides. Gartneri 2020. aasta uuringu kohaselt on andmejärv üks kolmest enim kasutatavast ärianalüütika tehnoloogiast [4]. Gartneri uuringus prognoositakse, et andmejärve kasutajate arv kasvab järgmise 12 kuu jooksul rohkem kui andmelao kasutajate hulk.

Järgnevas jaotises kirjeldatakse andmejärve tehnoloogilist olemust. Andmejärve arhitektuur annab hea ülevaate, kuidas andmejärv andmeid kasutab ning kuidas tehakse andmed erinevatele tarbijatele kättesaadavaks.

2.2.1 Andmejärve arhitektuur

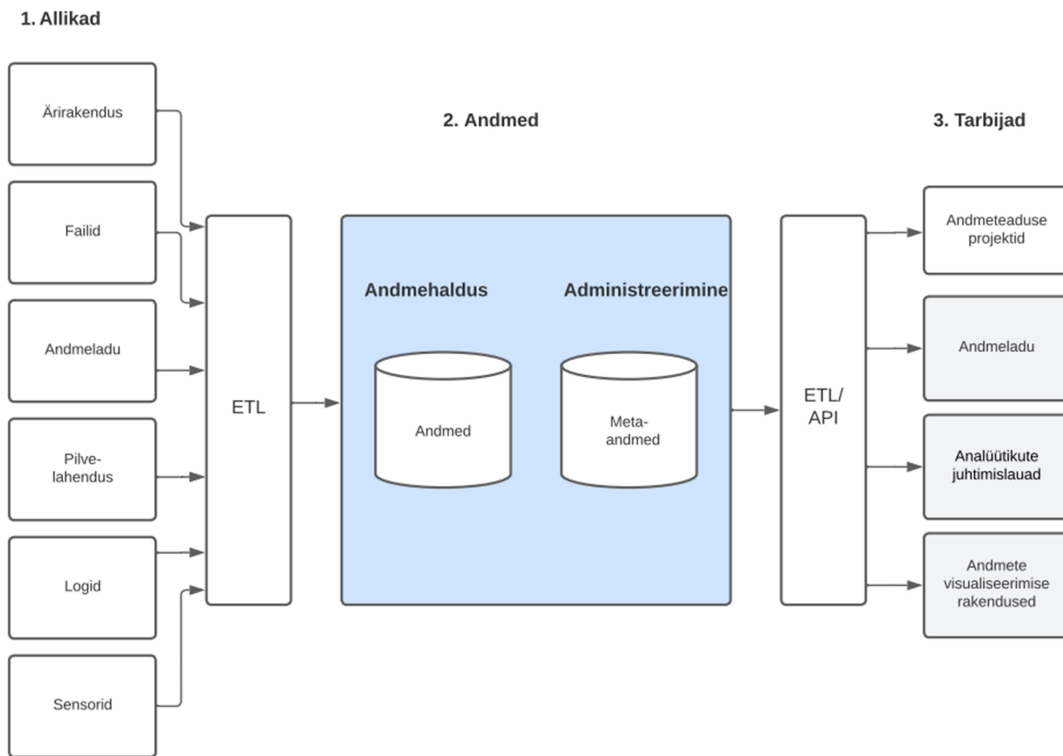
Andmejärve arhitektuuri osas levib erinevaid kontseptsioone. Tehnilised lahendused erinevad peamiselt selle poolest, kuidas andmejärv jagatakse erinevateks komponentideks ning kui palju ja millist andmete töötlemist andmejärves tehakse.

Sawadogo ning Darmont on oma töös võrrelnud erinevaid kasutusel olevaid arhitektuure. Autorid kategoriseerisid erinevad arhitektuurid tiigi (ingl *Pond*) ja tsooni (ingl *Zone*) arhitektuuriks [19].

Tiigi tüüpi arhitektuuris on andmejärv jaotunud eraldiseisvateks osadeks lähtuvalt andmete tüübist. Konkreetne tiik defineeritakse andmete tüübi järgi ning sellele vastavalt määratakse tiigile andmete salvestamise ja töötlemise vahendid. Üheks tiigiks võib olla struktureeritud andmete tiik ning seal kasutatakse andmete salvestamiseks ja töötlemiseks relatsioonilist andmebaasi. Teiseks tiigiks võib olla näiteks sensorandmete tiik ning seal kasutatakse andmete salvestamiseks ja töötlemiseks mõnda NoSQL andmebaasi (MongoDB [20], Redis [21]) või suurandmete töötlemise tarkvara (MapReduce [22], Apache Spark [23]). Kolmandaks tiigiks võib olla tekstiandmeid sisaldav tiik ning seal kasutatakse teksti töötlemise tarkvara (nt Apache OpenNLP [24], GATE [25]).

Tsoon tüüpi arhitektuuris on andmejärv jaotunud eraldiseisvateks osadeks lähtuvalt andmete puhastatuse astmest. Näiteks võib olla eraldi tsoon, mis tegeleb värskete andmetega, kus eelnevalt ei ole andmetele kontrolli või puhastusi rakendatud. Teiseks on tsoon, kus andmete puhastamisel on rakendatud tavapäraseid kvaliteedikontrolle. Kolmandaks on tsoon, kus asuvad usaldusväärsed andmed – andmed, mis on põhjalikult kontrollitud, puhastatud ja valmis kasutuseks näiteks analüüsitsoonis.

Üldistatult kirjeldab andmejärve arhitektuuri alljärgnev joonis 4 [26].



Joonis 4. Andmejärve arhitektuur

Arhitektuurijoonisest nähtub, et moodustub kolm peamist gruppi: allikad, andmed ning tarbijad. Allikatest liiguvad andmed andmejärve, kust tarbijad saavad vajalikke andmeid.

Vaatleme kolme gruppi ning neid esindavaid komponente lähemalt:

1. Andmeallikateks võivad olla andmed erinevatest lähteallikatest nagu majandustarkvara, andmeladu või pilvelahendused. Andmejärv võib kasutada ka seadmete poolt kogutud andmeid nagu kaamerate, sensorite andmed jt.

Andmed koondatakse kokku, rakendades ETL või ELT laadimisprotsessi. Kusjuures andmete teisendamist, mida teeb näiteks andmelao lahendus, klassikalise andmejärve kontseptsiooni kohaselt ei toimu. Andmed salvestatakse andmejärve lahenduses andmete algupärase olekus.

2. Andmete salvestus- ning halduskeskkond on disainitud toetamaks andmete turvalisust, skaleeritavust ning elastsust. Selles kihis toimub andmete ning neid kirjeldavate andmete e metaandmete salvestamine. Erinevate töövahendite abil võimaldatakse siin andmejärve andmed indekseerida ning rakendada vajalikud seadistused.

Andmejärve oluliseks komponendiks on metaandmed. Kuna andmejärve saabuvad andmed paljudest erinevatest lähteallikatest, siis on oluline, et andmete kasutajal oleks ülevaade, millised andmed andmejärves asuvad. Selleks on andmejärves alati metaandmete kataloog. See sisaldab järgmist infot [19]:

- milliseid andmeid andmejärv sisaldab ja kus (millistes kaustades, andmebaasides vms) need asuvad;
- millal on andmed andmejärve laaditud;
- millises formaadis andmed on (JSON, XML vms);
- milline on andmete struktuur ja tähendus (nt JSON failis kasutatavad atribuudid, atribuutide ärilised kirjeldused jms).

Andmete salvestamise vahendina kasutatakse andmejärves peamiselt kas andmebaasi (tiigi arhitektuuri korral) või jagatud failisüsteemi (tsooni arhitektuuri korral) [19]. Andmebaasideks on vastavalt andmetele, kas relatsiooniline andmebaas või NoSQL andmebaas. Võimalik on ka relatsioonilise ja NoSQL andmebaasi kombinatsioon ühes andmejärves.

2017. aasta uuringu kohaselt kasutati andmejärvedes ligi 75% ulatuses andmete salvestamiseks Apache Hadoop-i [27] jagatud failisüsteemi [28]. Apache Hadoop lahendus suudab käsitleda mistahes andmetüüpe, pakub andmete salvestamiseks kõrget tõrkekindlust ja hõlpsat laiendatavust. Siiski pole Hadoop üksinda kõikide andmestruktuuride käsitlemiseks efektiivne [28]. Lahenduseks oleks siinkohal Hadoop lahenduse kombineerimine kas relatsioonilise ja/või NoSQL andmebaasiga.

3. Andmejärve tarbijateks võivad olla andmeladu, andmete visualiseerimise rakendus (näiteks Tableau, Power BI jt), analüütilised rakendused, masinõppe mudelid ja vahendid (näiteks R programm). Joonisel halli taustaga markeeritud tarbijad on erijuhtumid, mida esineb andmejärve tarbijana vähesel määral. Peamisteks andmejärve tarbijateks on andmeteaduse projektid.

Andmejärve ja andmelao arhitektuuris leidub mitmeid sarnasusi. Tihti mõistetakse neid kahte lahendust kui alternatiivseid lahendusi [4], kuid mõlemad tehnoloogiad kannavad siiski erinevaid eesmärke ja rolle.

Andmejärv kogub kokku nii struktureeritud kui ka struktureerimata andmed. Vajalikke eeltegevusi andmetega, et valmistada andmed ette aruannete koostamiseks ning teostada nendel andmetel tuginevat analüüsi, andmejärve lahendus enamasti ei sisalda. Andmeladu

seevastu kogub kokku ennekõike struktureeritud andmed ja andmete eeltöötlus kuulub selle tehnoloogia põhifunktsionaalsuse hulka.

Andmeladu ja andmejärve võib kasutada koos. Lahendus, kus andmeladu ammutab andmeid andmejärvest, võimaldab rakendada nende kahe tehnoloogia kõige tugevamaid omadusi [19]. Andmejärv võimaldab hõlpsat ning odavat suure hulga andmete salvestamist, seega võib andmejärve kaaluda kui andmelao eellaadimisala või ODS-i [28]. Andmeladu korrastab ja eeltöötleb andmejärve andmed, et saaks need anda lõpptarbija kasutusse.

2.2.2 Andmejärve tugevused

Eelnevas jaotises kirjeldati lühidalt andmejärve üldist olemust ning arhitektuurilist ehitust. Terviklikuma ülevaate saamiseks käsitletakse käesolevas jaotises andmejärve tugevusi ning järgnevas peatükis andmejärve nõrkusi.

Andmejärvel on järgmised tugevad omadused [29]:

- Andmed on lihtsasti kättesaadavad igal ajahetkel. Seeläbi on võimalik andmeid analüüsida reaalaajas. Nimetatud võimalused on tingitud asjaolust, et üldjuhul andmejärves ei rakendata andmete eelnevat puhastamist ning teisendamist. Andmed on pärast importimist kohe kasutusvalmis.
- Suure hulga andmete kuluefektiivne salvestamine. Erinevalt näiteks andmelaost, ei vaja andmejärv kiireid, kuid samas kalleid SSD kõvakettaid. Salvestades andmed näiteks Apache Hadoop-i jagatud failisüsteemi, on kulud madalad tänu vabavarale ja soodsatele kõvaketastele.
- Andmejärv pakub suurt valikut eri variatsioonides andmeid (struktuur, formaat, päritolu jne).
- Andmete salvestamisel nende vormingut ei muudeta ning seeläbi on võimalik andmeid piiranguteta kasutada ja lisada.
- Andmed on rikkalikud ning mitmekesised, kuna andmed võivad olla pärit mistahes allikatest ja formaatides (sotsiaalmeedia, sõnumivahetused, fotod jne).
- Võimalus kasutada andmete uurimisel ning töötlemisel masinõppe erinevaid meetodeid. Andmejärv võimaldab salvestada erinevas formaadis andmeid ja väga suures mahus andmeid. Need on omadused, mis on väga vajalikud masinõppe rakendamisel.

Andmejärve eeliseid saab kõige efektiivsemalt rakendada just teadusprojektides, kus on konkreetne andmete kasutamise eesmärk ning ajaperiood. Andmejärve koondatud erineva formaadi ning kvaliteediga andmed on rikkalikuks sisendiks andmeteadlastele, kes saavad andmeid projekti vajadusest tulenevalt analüüsida ning vormindada.

2.2.3 Andmejärve nõrkused

Vastukaaluks eelmises peatükis vaadeldud andmejärve tugevustele, on loetletud ka peamised nõrkused. Andmejärvel on järgmised omadused, mida analüütikaprojekti tehnoloogia valimisel kindlasti peaks silmas pidama [19]:

- Andmete töötlemine ja analüüsimine vajab kõrge kvalifikatsiooniga andmeteadlast või spetsialisti (ei ole jõukohane vastava väljaõppeta inimesele).
- Andmete kooskõla ei ole garanteeritud, kuna andmed tulevad eri allikatest ning andmeid üldjuhul ei puhastata ega teisendada. Näiteks on võimalik olukord, kus üks ja sama andmeobjekt (näiteks klient) salvestatakse andmejärve erinevatest lähtesüsteemidest, erineva tõlgendusega (erinev kliendi mõiste) ning erineva semantikaga (ühel juhul identifitseerib kliendi tema ID, teisel juhul kliendi nimi).
- Andmejärvel puuduvad metoodilised ja tehnilised standardid. Kokkulepitud loogikate järgimine aitaks ära hoida andmejärve projekti ebaõnnestumise.
- Andmejärve tehnoloogia nõuab tõhusat metaandmete süsteemi andmete kasutamiseks ning andmete sisu õigeaks tõlgendamiseks. Metaandmete haldamiseks ebasobivate meetodite või tehnoloogiate kasutamisel võib andmejärv kergesti muutuda mittetoimivaks nõ andmesooks (ingl *Data Swamp*).

Andmejärve kui analüütikaprojekti tehnoloogia peamise kitsaskohana võib välja tuua asjaolu, et andmed ei ole ärianalüüsiks kohe kasutusvalmis. Erinevalt andmelaost ei ole ettevõtte juhul või analüütikul võimalik saada vajalikke andmeid enne, kui andmeteadlane või spetsialist on teinud andmejärve andmete osas põhjaliku puhastuse ja need ümberstruktureerinud.

Juhul, kui kokku kogutud andmetele ei ole rakendatud korrektset metaandmete süsteemi, mis võimaldab kasutajal andmeid õigesti tõlgendada ja seostada, võib andmejärv muutuda halvasti kasutatavaks. Korrektse metaandmete süsteemi korral on tuvastatav, mis tabelis andmed asuvad, millised on tabelite vahelised seosed, mis allikast on andmed pärit, millal andmed on laaditud jms. Korrektse metaandmete süsteemi korral on andmetel selge äriiline

tõlgendus: näiteks kauba maksumuse puhul on teada, kas hind sisaldab käibemaksu või mitte.

Mõnede allikate kohaselt sisaldab andmejärv mõningast andmete teisendamist ja eelnevat puhastamist. Samas tähendab see, et andmejärv muutub sarnasemaks andmelaole ning kaob andmejärve oluline omadus – hoiustada andmeid nende algupärasel kujul. Toorandmete ehk töötlemata andmete analüüsimine on eriti oluline ennekõike andmeteaduse projektides.

2.2.4 Andmejärve kokkuvõte

Andmejärv on uudne tehnoloogia ning esmakordselt mainitud aastal 2010. Andmejärv võimaldab koguda, säilitada ja töödelda mistahes andmeid ning tagada kasutajatele ligipääsu neile andmetele.

Käesolevas peatükis on vaadeldud erinevaid kasutusel olevad andmejärve arhitektuure ning esitatud üldine andmejärve arhitektuurijoonis koos selgitustega. Põhjalikumalt on arhitektuuri juures käsitletud andmete salvestamise viise. Sarnaselt andmelao peatükile, kirjeldati andmejärve tehnoloogia peamisi tugevusi ja nõrkusi.

Andmete perspektiivist vaadatuna saab andmejärvest tehtud ülevaate kokku võtta järgmiselt:

- andmejärves saab kasutada mistahes andmeid, nii struktureeritud kui ka struktureerimata andmeid;
- peamisteks andmejärve kasutajateks on andmeteadlased;
- andmed on andmejärve salvestatud jagatud failisüsteemi või andmebaasi (sõltuvalt valitud arhitektuurist);
- andmed liiguvad andmejärve läbi ETL/ ELT või API liidese.

Eelmises peatükis kirjeldatud andmeladu eristub andmejärvest mitmel moel. Tugevalt joonistub välja andmelao seos struktureeritud andmetega ning andmejärve seos mistahes struktuuriga andmetega. Andmeladu omab andmete eeltötluse ning puhastamise funktsiooni, kuid andmejärve klassikaline kontseptsioon seda funktsiooni ei kajasta. Siiski on mõlemale tehnoloogiale võrdselt omane koondada andmeid erinevatest allikatest, need hoiustada ning teha kättesaadavaks erinevatele kasutajagruppidele.

2.3 Loogiline andmeladu

Klassikaline andmeladu (ingl *Data Warehouse* – DWH) on ettevõtteid aastakümneid edukalt teeninud, kuid klassikalise andmelao arhitektuur ei ole alati parim lahendus. Aina kiiremini vajatakse uute aruannete valmimist, ettevõttesse lisandub järjest uusi andmete lähtesüsteeme ning nõudlus tehnoloogia paindlikkuse osas kasvab.

Lahendust pakub siinkohal loogiline andmeladu (ingl *Logical Data Warehouse* – LDW). Loogiline andmeladu on kaasaegne andmearhitektuur, mis võimaldab ettevõttel kasutada kõiki oma andmeid reaajas, olenemata sellest kus andmeid hoitakse, millises formaadis andmed on ning milliseid tehnoloogiaid andmete salvestamiseks ja neile juurdepääsuks kasutatakse [30].

Loogilise andmelao mõiste võeti aktiivsemalt kasutusele aastast 2011 [31]. Loogilise andmelao kontseptsioon kasvas välja klassikalisest andmelaost, ajendatuna soovist muuta andmeladu paindlikumaks ning rohkem vastavaks kaasaja vajadustele [31]. Mõistes esinev sõna „loogiline“ viitab siin asjaolule, et uus trend keskendub informatsiooni loogikale, mitte andmete hoiustamise mehhanismile [31].

Loogiline andmeladu on virtuaalne kiht, mis liidab klassikalise andmelao arhitektuuri füüsilised kihid üheks loogiliseks tervikuks [32]. Loogiline andmeladu suudab pärida andmeid reaajas ning erinevatest andmeallikatest, nii andmelaost kui ka teistest süsteemidest, sh suudab käsitleda ka erinevaid andmestruktuure.

Loogilise andmelao omadustest võib kõige enam kasu olla ettevõtetel, kellel on vajadus reaajas andmete järgi. Loogilise andmelao kasutajana võib välja tuua ettevõtte, kelle andmed asuvad füüsiliselt erinevates lähteallikates. Näiteks olukord, kus ettevõtte müügiandmed paiknevad nii ettevõtte e-poes (e-poe andmebaas), lokaalsetes müügiesindustes (majandustarkvara) kui ka erinevates MS Exceli failides, ettevõtte aga soovib täpselt näha, mitu toodet hetkeseisuga on müüdud, ning vajab selleks reaajas töötavat aruandlust.

Loogilise andmelao mõiste ei ole magistritöö autorile teadaolevalt eestikeelsesesse terminoloogiasse jõudnud. Samuti ei ole autoril andmeid, kas Eestis on loogilise andmelao arhitektuur mõnes ettevõttes kasutusel ning kas eksisteerib kohalik edasimüüja, kes pakub loogilise andmelao arhitektuuri toetavat platvormi.

Mujal maailmas on loogilise andmelao arhitektuur kasutusel mitmetes riikides. Suur hulk ettevõtteid kasutab ettevõtte Denodo loogilise andmelao arhitektuuri toetavat platvormi

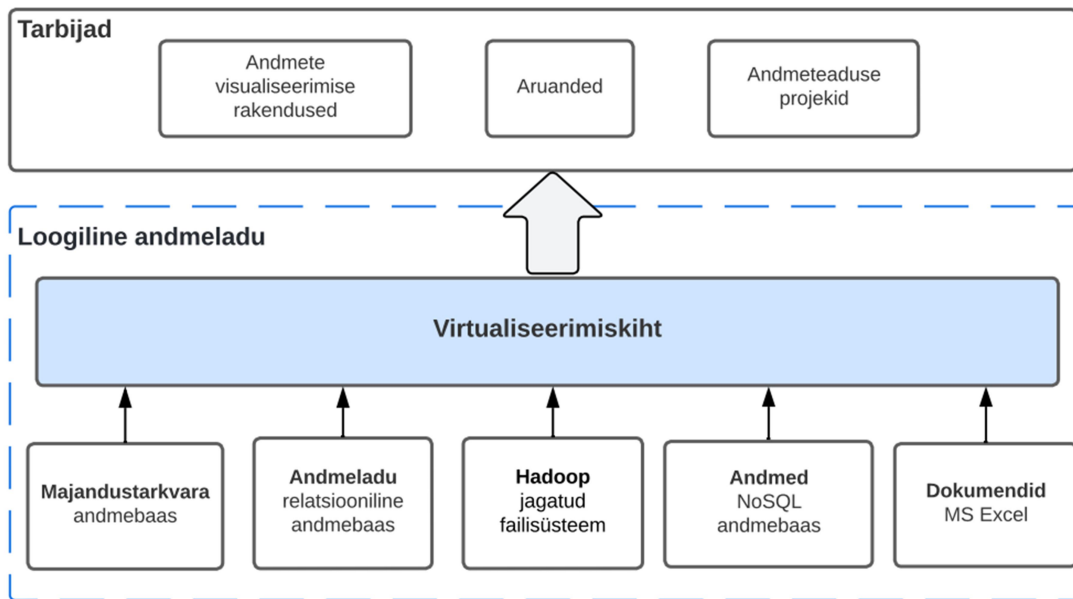
[33]. Denodo platvorm on kasutajate tagasisidena saanud 5 täрни süsteemis 4,5 täрни [34]. Tagasiside ning hinnangud on kasutajad andnud Gartneri poolt valideeritud tagasisidestamise keskkonnas. Lisaks tunnustas Gartner Denodo platvormi tiitliga „Klientide valik 2022“ andmete integratsiooni vahendite (ingl *Data Integration Tools*) kategoorias [34].

2.3.1 Loogilise andmelao arhitektuur

Eelmises jaotises kirjeldati loogilist andmeladu kui virtuaalset kihti, mis liidab klassikalise andmelao arhitektuuri füüsilised kihid üheks loogiliseks tervikuks. Loogilise andmelao arhitektuuri käsitlemiseks on mitmeid võimalusi. Üks võimalus loogilise andmelao arhitektuuri käsitlemiseks on jagada arhitektuur järgmistesse kihtidesse:

- Alumises arhitektuurikihis on riistvara (näiteks serverid) ning võrguühendus, mida nii klassikaline kui ka loogiline andmeladu tarbib.
- Järgmiseks kihiks on tarkvarakiht, mille peal töötavad erinevad haldussüsteemid nagu näiteks andmebaasihaldussüsteem.
- Seejärel on andmevarunduskiht, kuhu reeglina kuuluvad erinevad platvormid, kus andmeid hoiustatakse.
- Viimaseks kihiks on andmete arhitektuuri kiht. Selles kihis hallatakse kõiki andmetega seonduvaid võtmeküsimusi ning tagatakse andmetele ligipääs. Selles kihis asub loogiline andmeladu.

Kihtideks jagatud loogilise andmelao arhitektuuri kirjeldab joonis 5 [33]. Joonise selgitused ning loogilise andmelao lihtsustatud tööpõhimõte on lisatud joonise alla.



Joonis 5. Loogilise andmelao arhitektuur

Loogiline andmeladu on joonisel kujutatud sinise katkendliku piirjoonega. Loogilises andmelaos on kesksel kohal virtualiseerimiskiht, mis liidab temaga ühendatud lähteallikad üheks terviklikuks arhitektuuriks.

Lähteallikad omavad taristut (andmebaas, server, võrguühendus jm) ning oma arhitektuuri. Virtualiseerimiskiht omab loogikat ning teadmist, milliseid andmeid millistest lähteallikatest on vaja, milliseid reegleid andmete puhastamisel ning korrastamisel rakendada.

Lähteallikateks võivad loogilise andmelao korral olla erinevad ettevõttesisesed ja -välised infosüsteemid/andmebaasid (näiteks majandustarkvarad, kliendihaldusprogrammid jpt), veebilehed, sotsiaalmeedia ja partnerite süsteemid. Samuti võivad lähteallikateks olla failid ja dokumendid.

Virtualiseerimiskiht on mistahes asukohas paiknev tarkvara. Tarkvara olemuseks on teenus, mis tarbija tellimusel vahendab reaajas tarbija poolt küsitud andmeid. Virtualiseerimiskiht omab iga lähteallika andmemudelit ning reeglistikku, milliseid andmeid on tarvis pärida, kuidas lähteallika andmeid puhastada, korrastada ning milliseid tegevusi (näiteks arvutusalgoritmide rakendamine) on andmetele vaja rakendada.

Lihtsustatult võib öelda, et virtualiseerimiskiht on kombinatsioon ELT-st ning klassikalisest andmelao andmebaasist. Virtualiseerimiskiht omab kõiki ELT ning andmelao andmebaasi omadusi, välja arvatud andmete salvestamise funktsiooni. Sõltuvalt

edasimüüja tootest esineb mõningaid loogilise andmelao arhitektuure, mis sisaldavad andmete salvestamise funktsionaalsust.

Roll: Virtualiseerimiskiht algatab reaajas päringu vajalikesse lähteallikatesse kui tarbijalt saabub andmete tellimus ehk päring (näiteks vajutatakse Tableau müügiaruandes „Värskenda“ nuppu). Virtualiseerimiskiht koondab andmed kokku, rakendab vajalikud reeglid ning teostab varasemalt kirjeldatud protseduurid, et viia andmed tarbija jaoks vajalikule kujule. Seejärel saab tarbija ärirakendus päringule vastuse, mille sisuks on virtualiseerimiskihi poolt ette valmistatud andmed. Loogilise andmelao andmed on tarbija ärirakendusele kättesaadavad läbi kokkulepitud API liidese [35].

Loogilise andmelao tarbijad on peamiselt ärikasutajad (sarnaselt klassikalise andmelao tarbijatele). Ärikasutajatele teeb andmelaos ettevalmistatud andmed kättesaadavaks andmete visualiseerimise rakendus (näiteks Tableau, MS Power BI), juhtimislaud või mõni muu analüütikarakendus.

Joonisel 5 kirjeldatud loogilise andmelao arhitektuur on kujutatud alt üles põhimõttel, lähtudes eespool kirjeldatud kihtidest. Selgelt joonistub välja sarnasus varem vaadeldud andmejärve ning klassikalise andmelao tarbijate ning lähteallikate loetelu vahel. Lisaks ilmneb virtualiseerimiskihi rolli kirjelduses palju klassikalisel andmelao tööpõhimõtetega sarnaseid omadusi.

2.3.2 Loogilise andmelao tugevused

Loogilise andmelao eelised kattuvad suurel määral eelpool kirjeldatud klassikalise andmelao ning andmejärve eelistega. Loogiline andmeladu:

- võimaldab andmeid reaajas analüüsida;
- tagab korrastatud ja kvaliteetsed andmed;
- annab ligipääsu mitmekesistele ning rikkalikele andmetele, kuna andmed võivad olla pärit mistahes allikatest ja formaatides (sotsiaalmeedia, vestlused, fotod jne);
- suurendab andmeanalüüsi võimekust ja kiirust;
- tõstab ettevõtte müügitulu ja tugevdab ettevõtet läbi juhtimisotsuste, mis baseeruvad ajakohastel ning korrektsetel andmetel;
- on hõlpsalt laiendatav ning ettevõtte muutuvate ärivajaduste osas paindlik.

Võrreldes teiste analüütikalahenduste tehnoloogiatega, on loogilise andmelao suurimaks eeliseks paindlikkus erinevate andmete, lähteallikate ning tehnoloogiate suhtes. Ettevõtte

saab loogilise andmelao vahendusel andmeid vaadelda reaallajas ning andmed on eelnevalt kontrollitud ja korrastatud.

2.3.3 Loogilise andmelao nõrkused

Eelnevalt oli välja toodud loogilise andmelao omadus pakkuda reaallajas erinevatest allikatest pärit andmeid, mis on värsked ja mistahes formaadis. Seejuures peab loogiline andmeladu tagama ka kvaliteetsed ning korrastatud andmed. Äriprotsesside korral, kus on kõrged nõudmised, peab loogiline andmeladu arvestama ka lähteallikatega. Lähteallikatest andmeid pärides tuleb vältida lähteallikate ülekoormamist.

Suur osa ettevõtte lähteallikatest on operatiivinfosüsteemid (näiteks majandustarkvara), mis on töötajate poolt igapäevases kasutuses. Kui lähteallikad on päringute tõttu üle koormatud, muudab see lähteallika töö aeglaseks ning lähteallika töös võib esineda tõrkeid.

Loogilise andmelao ülesannete eduka täitmise eelduseks on kiire ja töökindel riistvara, tarkvara ning võrguühendus ja seda nii lähteallikate, tarbijate kui ka loogilise andmelao virtualiseerimiskihi osas. Suurte andmemahutuste ning aktiivsete päringute korral on kindlasti tegemist väga suure tehnoloogilise väljakutsega. Virtualiseerimiskiht peab suutma tagada andmed reaallajas kõigile tarbijatele, lähtesüsteeme seejuures koormamata.

Peamiseks loogilise andmelao kitsaskohaks on keerukas arhitektuur. Loogilist andmeladu kavandades peab arhitekt väga põhjalikult ja detailselt arvestama kõiki tehnilisi asjaolusid, sh nii tänaseid ärivajadusi ja olemasolevate süsteemide võimekust kui ka tuleviku suundi.

Loogilise andmelao kontseptsioon on väga võimekas, kuid samas seab riistvarale, tarkvarale ning võrguühendusele väga kõrged nõudmised.

2.3.4 Loogilise andmelao kokkuvõte

Loogiline andmeladu on üks uuemaid analüütikaprojektide tehnoloogiaid ning autori andmetel loogilise andmelao arhitektuuri Eestis veel kasutusel ei ole. Loogiline andmeladu vahendab andmeid reaallajas, olenemata lähteallikate hulgast, kasutusel olevatest tehnoloogiatest ning andmete formaadist.

Loogilise andmelao peatükis esitati lihtsustatud loogilise andmelao arhitektuuri joonis koos selgitustega. Põhjalikumalt vaadeldi loogilise andmelao virtualiseerimiskihti, mis omab kogu kontseptsioonis kesksel rolli.

Lisaks arhitektuurile toodi välja loogilise andmelao peamised tugevused ja nõrkused. Loogilise andmelao kontseptsioon on väga võimekas, kuid samas seab riistvarale, tarkvarale ning võrguühendusele kõrged nõudmised.

Andmete perspektiivist vaadatuna, saab loogilisest andmelaost tehtud ülevaate kokku võtta järgmiselt:

- loogilise andmelao vahendusel saab kasutada mistahes andmeid, nii struktureeritud kui ka struktureerimata andmeid;
- peamiseks loogilise andmelao kasutajateks on ärikasutajad, erinevad spetsialistid ning ettevõtte juhid;
- loogiline andmeladu vahendab andmed reaajas ja eelnevalt kokku kogutud ning korrastatud andmeid ei salvestata;
- andmed liiguvad loogilise andmelao lahenduses läbi päringute ning päringute tehnoloogia sõltub lähteallika või tarbija vajadustest.

Loogilise andmelao tugevuseks on andmete vahendamine reaajas, mis sarnaselt klassikalisele andmelaole sisaldab andmete puhastamist, korrastamist ning ärikasutajate jaoks ettevalmistamist. Nõrkusena tuleb märkida kõrgeid nõudeid riistvarale, tarkvarale ning võrguühendusele nende kiiruse ja töökindluse osas.

2.4 Andmejaotur

Eespool kirjeldatud analüütikalahenduste kõrval nagu andmeladu ja andmejärv on viimastel aastatel tärkamas uus tehnoloogia – andmejaotur (ingl Data Hub). Andmejaotur on alles arenemisjärgus ning kontseptsiooni kirjeldavaid allikaid napib. See omakorda on kaasa toonud, et tihti nähakse ekslikult andmejaoturit kui alternatiivset lahendust andmelaole või andmejärvele [4].

Eksiarvamuste vältimiseks ning selguse loomiseks on andmejaotur magistritöös eraldi välja toodud. Autorile teadaolevalt ei ole andmejaotur eestikeelsesesse terminoloogiasse ning kirjalikku käsitluse veel jõudnud.

Andmejaoturi sügavamaks mõistmiseks analüüsis töö autor andmejaoturit käsitlevaid allikaid. Materjale leidis alates aastast 2017 ning värskeim neist avaldati 2022. aastal. Autor tegi leitud artiklitest valimi, lähtudes allikate uudsusest ning põhjalikkusest. Artiklitel ([4], [36], [37], [38].) tuginev kokkuvõtlik analüüs ning autoripoolne tõlgendus andmejaoturi olemusest on esitatud järgnevatel lõikudes.

Andmejaotur võimaldab reaajas vahendada andmeid erinevate lähtesüsteemide (andmete haldussüsteemid) ning andmete tarbijate vahel. Lisaks võimaldab andmejaotur andmete kokku koondamist eri allikatest, andmete puhastamist ning ühtlustamist ning andmete kättesaadavaks tegemist tarbijatele. See tähendab, et olulise omadusena sisaldab andmejaotur ka andmehaldust (ingl *Data Management*), mis klassikaliselt sisaldub ETL/ELT lahendustes.

Andmejaotur pakub ligipääsu kõigile andmetarbijatele, olgu siis tegemist tarkvaralise rakenduse, andmeteadlase või ärikasutajaga. Ettevõtte erinevad ärirakendused saavad andmeid andmejaoturi vahendusel, kasutades sobivat API liidest.

Erinevalt klassikalisest andmelaost ning loogilisest andmelaost ei keskendu andmejaotur andmete analüütilisele kasutamisele (näiteks arvutuslike algoritmide rakendamine andmetele, agregeeritud andmete ja aruannete loomine). Lisaks ei tegele andmejaotur üldjuhul andmete salvestamisega, kuid sõltuvalt andmejaoturi edasimüüjast võib vastav tarkvara sisaldada ka teatud määral andmete salvestamise võimekust.

Andmejaoturi teenus keskendub konkreetsele ja piiritletud andmehulgale ning iga andmehulga vahendamiseks on eraldiseisev andmejaoturi teenus. Näiteks on üks teenus mõeldud kliendi andmete kokku koondamiseks ja vahendamiseks, teine teenus toote andmete kokku koondamiseks ja vahendamiseks jne.

Kasutatud allikatest leitud andmejaoturi kirjeldusele tuginedes võib järeldada, et andmejaotur ei ole terviklik analüütikalahendus, nagu seda näiteks on andmeladu ja andmejärv. Andmejaotur on pigem üksik analüütikalahenduse komponent, olles piiratud hulga andmete vahendaja.

Eestis on andmejaotur uudne. Internetiotsingu tulemusel on tuvastatav Eestis vaid üks *Data Hub*⁵, milleks on ESTHub. ESTHub-i on kirjeldatud järgnevalt [39]:

ESTHub on riiklik satelliidiandmete keskus, mis võimaldab andmete otsimist ja allalaadimist läbi spetsiaalse portaali. Lisaks on kasutajatel võimalik töödelda andmeid omaenda algoritme kasutades. ESTHub vahendab kasutajatele ainult Copernicuse programmi andmeid.

Mujal maailmas on andmejaotur kasutusel mitmetes riikides ning leidub mitmeid andmejaoturi edasimüüjaid. Andmejaoturi tarkvara edasimüüjateks on näiteks Informatica (Data Integration Hub nimetuse all) [40], Cloudera (Data Hub nimetuse all) [41] ning

⁵Andmejaoturi mõistet Eesti kirjalikes allikates veel ei esine.

Semarchy (Semarchy Unified Data Platform nimetuse all) [42]. Kõik nimetatud edasimüüjad pakuvad andmejaoturi tarkvara teatud erinevustega ning lisavõimalustega.

Andmejaotur võib olla heaks valikuks kui olemasolev andmehalduse lahendus ei võimalda paindlikkust ning laienemist. Samuti on andmejaotur abiks olukordades, kus andmetele on vaja rakendada täiendavat või uut andmehaldust. Andmejaoturit on võimalik rakendada ka andmelao ja andmejärve lahendustes, kus andmejaotur pakub andmejärvele või andmelaole juba eelnevalt kokku kogutud andmeid. Seeläbi väheneb andmelao/andmejärve koormus, kiirenevad analüütikalahenduse andmevahetuse protsessid ning suureneb analüütikalahenduse paindlikkus ning laiendatavus.

Kuna andmejaoturi näol ei ole tegemist analüütikavahendiga, ei ole käesoleva töö raames seatud eesmärgiks uurida ning kirjeldada andmejaoturit sama sügavuti, kui seda tehti klassikalise ja loogilise andmelao ning andmejärve korral. Lisaks on andmejaotur sedavõrd uudne, et puuduvad piisavad allikad, mille põhjal uurimist läbi viia.

Leitud materjalide põhjal saab kokkuvõtvalt väita, et andmejaoturi näol on tegemist tulevikutrende arvestades võimeka andmete vahendamise ning haldamise tehnoloogiaga. Andmejaotur on paindliku arhitektuuriga ning tema funktsionaalsus sõltub ennekõike konkreetsest andmejaoturi edasimüüja tarkvarast või ettevõtte ärivajadusest, kuna andmejaoturit on võimalik realiseerida ka arendusprojektina.

3. Käsitletud tehnoloogiate võrdlus

Käesolevas magistritöös on käsitletud nelja tehnoloogiat: andmeladu, andmejärv, loogiline andmeladu ja andmejaotur.

Nimetatud tehnoloogiate analüüsil on käesolevas töös lähtutud järgmistest küsimustest:

- 1) millised on lahenduses kasutatavad andmed;
- 2) kes ja milleks neid andmeid kasutab;
- 3) kuidas andmeid talletatakse;
- 4) kuidas andmeid analüütikalahendusse ja selle sees liigutatakse.

Tuginedes analüüsi tulemustele on koostatud tabel 1, mis koondab info kõikidest siin töös käsitletud tehnoloogiatest.

Tabel 1. Andmelao, andmejärve, loogilise andmelao ning andmejaoturi omaduste võrdlus

	Andmeladu	Andmejärv	Loogiline andmeladu	Andmejaotur
Andmete formaat	Peamiselt struktureeritud	Struktureeritud ja struktureerimata	Struktureeritud ja struktureerimata	Struktureeritud ja struktureerimata
Peamised tarbijad	Ärikasutajad	Andmeteadlased	Ärikasutajad	Ettevõtte erinevad ärirakendused
Andmete hoiustamine	Jah	Jah	Ei	Enamasti ei
Andmevahetus	ELT/ETL	ELT/ETL/API	API	API
Andmete puhastamine ja standardiseerimine	Jah	Ei	Jah	Jah
Andmete ettevalmistamine aruandluseks	Jah	Ei	Jah	Ei
Reaalajas andmed	Ei	Jah	Jah	Jah

Tabelist 1 nähtub, et andmeladu, loogiline andmeladu, andmejärv ning andmejaotur ei ole teineteisele alternatiivsed tehnoloogiad. Ühegi tehnoloogia korral ei esine tabelis samade väärtustega rida. Enim erinevusi leidub andmete aruandluseks ettevalmistamise ning andmete hoiustamise osas.

Suurimaks sarnasuseks on loogilise andmelao, andmejärve ning andmejaoturi omadus pakkuda andmeid reaalajas, samuti on võimalus hoiustada/vahendada kõiki tüüpi andmeid (struktureeritud ja struktureerimata andmeid).

Magistritöös käsitletud tehnoloogiad ei ole teineteise suhtes alternatiivsed lahendused, kuid teades nende omadusi, saab neid edukalt omavahel kombineerida. Seda saab teha nii täiesti uute analüütikaprojektide arendamisel kui ka olemasolevate analüütikaprojektide täiendamiseks ja kaasajastamiseks.

Sellisel on võimalik ühe tehnoloogia tugevusega katta teise tehnoloogia kitsaskohta. Samuti saab tehnoloogiaid koos kasutada, lähtudes vaid nende funktsionaalsest rollist. Näitena võib siin tuua loogilise andmelao arhitektuuri, mille lähteallikateks on klassikaline andmeladu ning andmejärv. Lisaks võib sellise lahenduse arhitektuuri kuuluda ka andmejaotur, mis klassikalist andmeladu ja/või andmejärve värskete andmetega varustab.

Andmelaol, loogilise andmelaol, andmejärvel ning andmejaoturil on erinevad omadused, kuid nad omavad ühist eesmärki: tagada õigel hetkel õiged andmed õigetele tarbijatele. Küsimus on vaid ettevõtte andmetes ning ärivajadustes.

Kokkuvõte

Magistritöös analüüsiti ja kirjeldati analüütikaprojektide realiseerimise tehnoloogiaid suurte analüütikaprojektide vaatest. Töös anti ülevaade tehnoloogiate arhitektuurist, tugevustest, nõrkustest ning võrreldi käsitletud tehnoloogiate omadusi omavahel.

Magistritöö üheks eesmärgiks oli tõestada hüpotees, et eestikeelses teaduskirjanduses ei ole varem vaadeldud analüütikalahendusi laiemas vaates, kus oleks käsitletud erinevaid (sh ka kõige uuemaid) tehnoloogiaid koos nende võrdlusega. Hüpotees sai tõestatud läbi teaduskirjanduse analüüsi. Üheski leitud allikas ei leidunud käesoleva magistritööga sarnase eesmärgipüstisega käsitlust, kus lugejal oleks võimalik saada infot, milliseid erinevaid (kaasaegseid) vahendeid on ärianalüüsiks võimalik kasutada.

Analüüsi käigus leitud kirjalikud allikad käsitlesid ennekõike kas konkreetset tehnoloogilist vahendit (nagu näiteks Hadoop, ELT protsess) või analüütikalahendust konkreetse ettevõtte või projekti näitel, mis ei võimaldanud saada laiemat, valdkonnaalast ülevaadet ja tehnoloogia arhitektuuri laiemat käsitlust. Uuemaid analüütikaprojektide realiseerimise tehnoloogiaid käsitleti vaid ühes lõputöös (andmejärve näitel).

Käesoleva magistritöö teine eesmärk oli luua eestikeelne allikas, mis kirjeldaks uuemaid analüütikalahenduste tehnoloogiaid. Tulenevalt eesmärgist viidi läbi järgmised tegevused:

- uuriti, millised on peamised analüütikaprojektide realiseerimise tehnoloogiad, mida kasutatakse suuremate projektide korral;
- kirjeldati uurimise käigus leitud tehnoloogiaid, sh uusimaid tehnoloogiaid;
- kirjeldati uurimise käigus leitud tehnoloogiate omavahelisi erinevusi.

Loetletud tegevuste tulemid on kirjeldatud käesoleva magistritöö teises peatükis ning on magistritöö avaldamisel eestikeelseks allikaks, mis kirjeldab uuemaid analüütikalahenduste tehnoloogiaid.

Tehtud magistritöö pakub edasiarendamise võimalusi uuteks lõputöö temadeks. Töös kirjeldatud uusimad analüütikaprojektide realiseerimise tehnoloogiad vajavad igaüks põhjalikumat ning tehnilisemat laadi uurimist ning käsitlemist, mida käesolevas töös magistritööle seatud mahupiirangu tõttu ei olnud võimalik teostada.

Magistritöö lõpptulemust saavad praktilise sisendina kasutada Eesti analüütikavaldkonna tellijad (ettevõtted, kes soovivad uut analüütikalahendust või olemasoleva kaasajastamist)

ja arendajad (ettevõtted, kes pakuvad tellijatele ärivajadustele vastavaid analüütikalahendusi). Ettevõtted saavad magistritööst esialgse teadmise analüütikavaldkonnast, sh info valdkonna uusimatest tehnoloogiatest. Lisaks annab magistritöö vajaliku sisendi, et arendajad saaksid pakkuda paremaid lahendusi ning tellijal oleks võimalus valida sobivaim lahendus.

Viidatud kirjandus

- [1] Techopedia. Dictionary. <https://www.techopedia.com/dictionary> (04.03.2022)
- [2] Gartner. Glossary. <https://www.gartner.com/en/information-technology/glossary/> (03.03.2022)
- [3] Negash S., Gray, P. Business Intelligence. AMCIS 2003 Proceedings. 2003.
- [4] Goasduff L. The Best Ways to Organize Your Data Structures. Gartner, 2020. <https://www.gartner.com/smarterwithgartner/the-best-ways-to-organize-your-data-structures> (22.02. 2022)
- [5] Microsoft Power BI. <https://powerbi.microsoft.com/en-us/> (04.04.2022)
- [6] Tableau. <https://www.tableau.com/> (04.04.2022)
- [7] Gartner. Top 10 Data and Analytics Trends for 2021. <https://www.gartner.com/smarterwithgartner/gartner-top-10-data-and-analytics-trends-for-2021> (05.04.2022)
- [8] Gartner. About. <https://www.gartner.com/en/about> (03.03.2022)
- [9] Rääbis A. Andmeaitade loomise koolitusseminar, 2022. <https://www.kratid.ee/andmeteadus> (25.02.2022)
- [10] Lisowski E. Data Warehouse Implementation. Addepto, 2021. <https://addepto.com/implement-data-warehouse-business-intelligence/>(06.04.2022)
- [11] The Global Data Management Community. Inmon Bill William. <https://www.dama.org/cpages/william-bill-inmon> (19.03.2022)
- [12] Inmon, W. H. Building the Data Warehouse. Second Edition. New York: John Wiley & Sons.1996.
- [13] TEHIK. Info tervishoiutöötajale. <https://www.tehik.ee/info-tervishoiutootajale> (25.02.2022)
- [14] Kratt. Töötukassa otsustustugi OTT. <https://www.kratid.ee/t%C3%B6%C3%B6tukassa-otsustustugi-ott> (25.02.2022)
- [15]. Resta OÜ. Kliendid. <http://resta.ee/kliendid/> (25.02.2022)
- [16] Maguire J. Top 10 Benefits of Data Warehouse. Datamation, 2020. <https://www.datamation.com/big-data/top-10-benefits-of-a-data-warehouse/> (19.02.2022)
- [17] Business Intelligence Data Warehousing. Data Warehouse Benefits. <http://www.bi-dw.info/data-warehouse-benefits.htm> (09.04.2022)
- [18] Dixon, J. James Dixon's Blog. Pentaho, Hadoop, and Data Lakes, 2010. <https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/> (10.02.2022)
- [19] Sawdago P., Darmont J. On Data Lake Architectures and Metadata Management. *Journal of Intelligent Information Systems*, 2020. No 56, p. 97-120.
- [20] MongoDB. <https://www.mongodb.com> (25.04.2022)
- [21] Redis. <https://redis.io> (25.04.2022)

- [22] HADOOP. MapReduce Tutorial. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html (25.04.2022)
- [23] Apache Spark. <https://spark.apache.org> (25.04.2022)
- [24] Apache OpenNLP. <https://opennlp.apache.org> (25.04.2022)
- [25] GATE. <https://gate.ac.uk> (25.04.2022)
- [26] Pedamkar P. Data Lake Architecture. EDUCBA. <https://www.educba.com/data-lake-architecture/> (10.02.2022)
- [27] Apache Hadoop. <https://hadoop.apache.org/> (15.04.2022)
- [28] Russom P. Data lakes purposes, Practices, Patterns and Platforms. TDWI research. 2017.
- [29] Lamiya F. R. IT Explained: What Is a Data Lake. TechAcute, 2021. <https://techacute.com/data-lake/> (12.02.2022)
- [30] Moxon P. Logical Data Warehouses and Why They Matter. Data Virtualization, 2016. <https://www.datavirtualizationblog.com/logical-data-warehouses-matter/> (10.04.2022)
- [31] Merv A. Mark Beyer, Father of the Logical Data Warehouse, Guest Post. Gartner Blog, 2011. <https://blogs.gartner.com/merv-adrian/2011/11/03/mark-beyer-father-of-the-logical-data-warehouse-guest-post> (10.04.2022)
- [32] Russom P. Defining the Logical Data Warehouse. TDWI, 2015. <https://tdwi.org/Articles/2015/10/20/Defining-the-Logical-Data-Warehouse> (11.04.2022)
- [33] Denodo. Data Virtualization for Logical Data Warehouse. <https://www.denodo.com/en/solutions/horizontal-solutions/logical-data-warehouse> (21.04.2022)
- [34] Gartner Peer Insights. Denodo Platform Reviews. <https://www.gartner.com/reviews/market/data-integration-tools/vendor/denodo/product/denodo-platform> (21.04.2022)
- [35] Dilmegani C. What is Data Virtualization? Benefits, Case Studies & Top Tools. AI Multiple, 2020. <https://research.aimultiple.com/category/data/data-virtualization/> (25.04.2022)
- [36] Wells D. Data Hubs – What’s Next in Data Architecture. Eckerson Group, 2019. <https://www.eckerson.com/articles/data-hubs-what-s-next-in-data-architecture> (17.04.2022)
- [37] Semarchy. Why the Data Hub is the Future of Data Management, 2022. <https://www.semarchy.com/blog/why-the-data-hub-is-the-future-of-data-management/> (17.04.2022)
- [38] Semarchy. How to differentiate a Data Hub, a Data Lake and a Data Warehouse, 2020. <https://www.semarchy.com/blog/how-to-differentiate-a-data-hub-a-data-lake-and-a-data-warehouse/> (25.04.2022)
- [39] CGI. Eesti riikliku sateliidiandmekeskuse ESTHub analüüs. CGI. 2020.
- [40] Informatica. Data Integration Hub. <https://www.informatica.com/products/data-integration/integration-hub/data-integration-hub.html> (18.04.2022)

- [41] Cloudera. Data Hub. <https://www.cloudera.com/products/data-hub.html> (18.04.2022)
- [42] Semarchy. The Semarchy Unified Data Platform. <https://www.semarchy.com/platform/> (18.04.2022)

Lisa 1. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Agne Rääbis,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

„Analüütikaprojektide realiseerimise tehnoloogilised lahendused“,

mille juhendajateks on Anne Villemis ja Ain Rääbis,
reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi
DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Agne Rääbis

07.05.2022