TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Informaatikainstituut

IDK40LT

Enelin Kavak 140143

ANDMEKVALITEEDI PROBLEEMIDE LAHENDAMINE ANDMEAIDAS

Bakalaureusetöö

Juhendajad: Jekaterina Tsukrejeva
magistrikraad
õppejõu assistent;
Silver Saar
bakalaureusekraad
vanem testianalüütik

Tallinn 2016

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Enelin Kavak

23.05.2016

Annotatsioon

Käesoleva bakalaureusetöö eesmärgiks on analüüsida erinevad andmekvaliteediga seotud probleeme andmeaidas ning leida võimalusi nende ennetamiseks või lahendamiseks. Töö autor keskendub andmete profileerimise etapil esinevate andmekvaliteedi probleemide analüüsile. Töö tulemusena saadud lahendusi on võimalik kasutada reaalsetes andmete profileerimise projektides, et vältida peamisi tekkivaid probleeme.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 37 leheküljel, 3 peatükki, 5 joonist, 3 tabelit.

Abstract

Solving Data Quality Problems in Data Warehouse

The aim of this Bachelor's thesis is to analyze different data quality problems in data warehouse and to find solutions or ways to prevent those problems. The author focuses on problems that occur in data integration and data profiling phase. The results of this thesis can be used in real life data profiling projects, in order to prevent different problems.

The thesis is in Estonian and contains 37 pages of text, 3 chapters, 5 figures, 3 tables.

Lühendite ja mõistete sõnastik

Andmete teisenduskiht	*Staging Area*. Kiht andmeallikate ja andmeaida või andmevaka vahel, mida kasutatakse andmete hoiustamiseks. Teisenduskiht on tavaliselt ajutine ja selle sisu kustutatakse pärast andmete laadimist sihtkohta.
Andmevakk	*Data Mart*. Juurdepääsu kiht andmelao keskkonda, mida kasutavad lõppkasutajad andmete saamiseks. Andmevakk on andmeaida alamhulk, mis on tavaliselt suunatud konkreetsele ärivaldkonnale või meeskonnale.
BI	*Business Intelligence.* Äriteave – igasugune organisatsiooni ajalugu, jooksvat olukorda ja tulevikuperspektiive puudutav teave.
DWH	*Data Warehouse*. Andmeait – tööotstarbeline ühiskasutusega andmekogu, võib koosneda mitmest mistahes tüüpi andmeid sisaldavast andmebaasist, hõlmates organisatsiooni kõiki andmeressursse.
ETL	*Extract-Transform-Load.* Eraldamine-tranformeerimine-laadimine. Kolm funktsiooni, mis on kombineeritud üheks tööriistaks eesmärgiga võtta andmed ühest andmebaasist ja laadida teise andmebaasi.
Vastavusdokument	*Mapping document*. Dokument, mis sisaldab endas allika ning sihtkoha ärireeglite informatsiooni. Dokument leiab suurimat rakendust ETL protsessi arendaja poolt, kes disainib ja arendab ETL protsesse.

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Ettevõtted koguvad erinevaid andmeid, mis on vajalikud ettevõtte efektiivseks toimimiseks. Kogutavate andmete hulka kuuluvad näiteks andmed ettevõtte klientide, töötajate ning pakutavate teenuste ja toodete kohta. Kui andmehulgad on suured, siis selleks, et andmetest vajalikku informatsiooni kätte saada, tuleb andmed loogiliselt struktureerida. Selle jaoks kasutatakse andmeaita. Selleks, et informatsioon oleks ka täpne ja kasutatav, peab andmekvaliteet olema hea. Erinevad protsessid, mis andmeid töötlevad, mõjutavad andmete kvaliteeti kas positiivselt või negatiivselt. Selleks, et välja selgitada, kas andmed vastavad ettevõtte poolt seatud reeglitele ja nõuetele, kasutatakse andmete profileerimist. Andmete profileerimise abiga saab tuvastada erinevaid andmekvaliteedi probleeme ning neid lahendades parandada andmekvaliteeti. Töö autor soovib rõhutada, et andmete profileerimine ei paranda andmekvaliteeti vaid võimaldab andmekvaliteeti hinnata.

1.1 Eesmärk

Käesoleva töö eesmärgiks on analüüsida ja leida võimalikke lahendusi erinevatele andmekvaliteedi probleemidele andmete profileerimise etapis andmeaidas. Töös lahendatavad probleemid pärinevad R. Singh ja Dr. K. Singh koostatud uurimustööst „*A Descriptive Classification of Causes of Data Quality Problems in Data Warehouse*“ ehk „Kirjeldav klassifikatsioon andmekvaliteedi probleemide põhjustest andmeaidas“. Käesolev bakalaureusetöö on eelkõige mõeldud kasutamiseks andmete profileerimise projektides ning tulemusena saadud lahenduste kasutamine aitab ennetada erinevaid negatiivseid üllatusi andmekvaliteedi kohta projekti lõppfaasis.

Töö autor leidis kirjandusega tutvumise käigus, et varasemalt on erinevatele probleemidele häid lahendusi pakutud D. Vucevic ja W. Yaddow koostatud raamatus „*Testing the Data Warehouse Practicum*“ ehk „Andmeaida testimise praktikum“. Siiski ei käsitletud antud raamatus kõiki selle töö aluseks olevas uurimustöös välja toodud probleeme.

1.2 Metoodika

Kasutatav metoodika hõlmab igas alampeatükis välja toodud probleemide lahendamise või ennetamise teoreetilist analüüsi. Antud töö põhineb erinevatele teadusartiklitele ja raamatutele, mida töö autor uuris eesmärgiga leida parimad võimalused andmekvaliteedi probleemide lahendamiseks või ennetamiseks. Kuna töö autor töötab testianalüütikuna finantsettevõttes, millel on üks suurimaid andmeaitasid Eestis, siis on probleemide lahenduste otsimisel lähtutud just testianalüütiku igapäevatööst ja arvesse võetud ka töökogemust.

1.3 Ülevaade tööst

Antud töö koosneb kahest osast. Töö esimeses osas antakse ülevaade erinevatest baasmõistetest nagu andmeait, ETL protsessid ja äriteave. Lisaks kirjeldatakse ka andmekvaliteedi olulisust andmeidas ning mida kujutab endast andmete profileerimine.

Teises osas keskendutakse võimalikele andmeaida profileerimise etapis tekkivatele probleemidele ning pakutakse võimalusi väljatoodud probleemide ennetamiseks ja lahendamiseks.

2 Andmeait

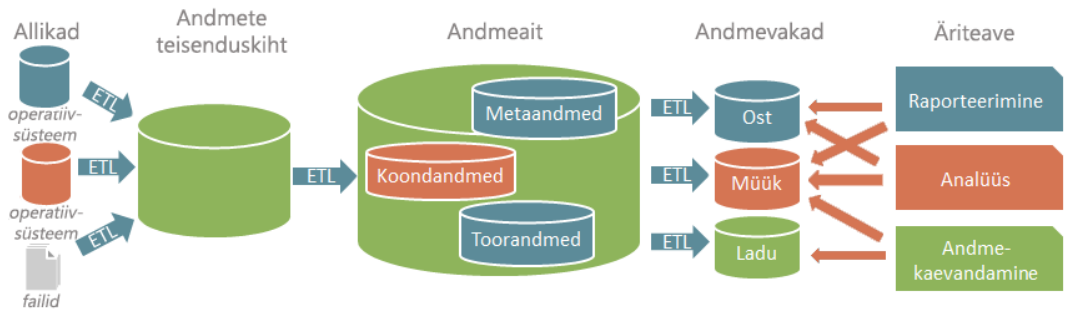
Iga ettevõtte jaoks on oluline olla edukas. Edukaks saamise võtmeks on kasulikud otsused, mida tehakse kaalutletult ja põhjendatult. Ettevõtte jaoks kasulike otsuste vastuvõtmisel on abiks andmeaida kasutamise. Selles peatükis kirjeldatakse andmeaita ja selle tähtsamaid osi.

Andmeaidal on erinevaid definitsioone. Infotehnoloogia (edaspidi IT) ja sidetehnika seletava sõnaraamatu e-teatmik defineerib andmeaita kui tööotstarbelist ühiskasutusega andmekogu, mis võib koosneda mitmest mistahes tüüpi andmeid sisaldavast andmebaasist, hõlmates organisatsiooni kõiki andmeressursse [1]. Kõige tuntum andmeaida definitsioon pärineb aga Bill Inmon'ilt, kes on tuntud ka kui andmaida isa: „Andmeait on subjekt-orienteeritud, integreeritud, ajast sõltuv ja püsiv andmete kogumik, mis toetab juhtkonda otsuste tegemisel" [2]. Need neli põhilist andmeaida omadust on [3]:

- subjekt-orienteeritus – andmed klassifitseeritakse suuremate valdkondade järgi, mis on eelnevalt andmemudelis defineeritud;
- integreeritus – andmeid kogutakse mitmetest allikatest ning liidetakse ühtseks tervikuks;
- ajast sõltuvus – kõik andmed on seostatud kindla ajaperioodiga;
- püsivus – andmeid ei hävitata kunagi: kui andmeaita kirjutatakse rida andmeid, siis seda rida enam ei muudeta.

2.1 Arhitektuur

Joonis 1 on toodud üldine andmeaida arhitektuur, kuid andmeaida peaks siiski kohandama vastavalt ettevõtte struktuurile ja vajadustele.



Joonis 1. Andmeaida üldine arhitektuur [4].

Eraldamine-tranformatsioon-laadimine (edaspidi ETL) on protsess, mille abil koondatakse kõigi organisatsiooni osakondade vahel hajutatud andmed ühtseks andmeaidaks [3]. Andmeaida arhitektuur koosneb järgmistest osadest [4]:

- andmeallikad, mis varustavad andmeaita andmetega. Joonis 1 on toodud allikatena operatiivsüsteemid ja failid (näiteks Exceli failid, xml-formaadis failid);
- ETL protsessid, mille abil laetakse andmeid ühest asukohast teise;
- andmete teisenduskiht, mida kasutavad ETL protsessid andmete laadimisel;
- andmeait ise, mis sisaldab metaandmeid, koondandmeid ja toorandmeid, mis on kantud sinna ETL protsesside abil andmete teisenduskihist;
- andmevakad, kuhu laetakse ETL protsesside abil transformeeritud andmeid;
- äriteave, mis saadakse andmevaka andmete analüüsist ja andmekaevandamisest ning mille põhjal tehakse ka raporteerimist.

2.2 ETL protsessid

ETL protsesside abil koondatakse kõigi organisatsiooni osakondade vahel hajutatud andmed ühtseks andmeaidaks [3]. Joonis 2. on kujutatud ETL protsesside kasutamine.

Joonis 2. ETL protsesside kasutamine [5].

ETL protsessid koosnevad järgmisest etappidest [3]:

1. Eraldamine – etapi eesmärgiks on eraldada allikast vastavate kriteeriumite või filtrite põhjal vajalikud andmed ning valmistada need ette transformatsiooni etapi jaoks.

2. Transformatsioon – etapi eesmärgiks on allikast eraldatud andmete transformeerimine vastavalt struktuurile, mis on sobivaim andmeaidas kasutamiseks. Transformatsiooni etapp standardiseerib, integreerib, puhastab, täiendab ja koondab andmeid ning loob andmekogud laadimise protsessi jaoks. Andmete transformatsioon on andmeaida juures üks suurim ja raskeim väljakutse.

3. Laadimine – etapi eesmärgiks on laadida eelnevalt eraldatud ja transformeeritud andmed struktureeritult sihtkohta, milleks võib olla andmete teisenduskiht, andmeait või andmevakk.

2.3 Äriteave

Ettevõtted koguvad oma klientide kohta suurel hulgal erinevaid andmeid, kuid selleks, et kogutud andmed ka ettevõttele kasulikuks osutuksid, tuleb rakendada äriteabe kontseptsiooni. Äriteave on tehnoloogiapõhine protsess andmete analüüsimiseks ja praktilise teabe esitamiseks ettevõtete juhtidele ja teistele lõppkasutajatele, et neid aidata teadlike otsuste tegemisel [6]. Äriteabe abil saadud informatsioon võimaldab tuvastada nii kliendi vajadusi kui ka pakkuda ülevaadet kliendi otsuste tegemise protsessist, ettevõtte tegevusvaldkonna trendidest, üldistest majanduslikest trendidest ja majanduslikust seisust, sotsiaalsetest trendidest ning teistest huvivaldkondadest [3].

Peamine äriteabe osa on andmete analüüs – andmekaevandamine, ennustav analüüs, statistiline analüüs ja suurandmete analüüs [6].

2.4 Andmekvaliteet andmeaidas

Õigeid äriotsuseid on keeruline teha ebakorrektsete andmete põhjal, seega on kvaliteetsed andmed andmeaida alustalaks. Ka andmeaitade puhul kehtib IT valdkonnas tuntud akronüüm GIGO (ingl. k *Garbage In, Garbage Out*), mis tähendab, et vigaste sisendandmete põhjal ei saa andmeait pakkuda kasutajale kvaliteetseid väljundandmeid. Andmeid mõjutavad mitmed protsessid, mis toovad andmeid andmeaida keskkonda. Andmete kvaliteet võib olla ohustatud sõltuvalt sellest, kust andmed saadakse ning kuidas andmed sisestatakse, integreeritakse, säilitatakse, töödeldakse (ETL protsessid) ja andmeaita laetakse. Hoolimata pingutustest, eksisteerib igas andmeaidas siiski mingi hulk vigaseid andmeid [7].

Kõige suurema osa andmekvaliteedi tagamisest moodustab andmete profileerimine [3].

2.4.1 Andmete profileerimine

Andmete profileerimine on kindla andmeallika süstemaatiline analüüs andmeaida jaoks, et mõista andmete struktuuri, sisu, suhteid ning tuletamise reegleid [3]. Andmeaida puhul tuleks kaaluda andmete profileerimist:

- andmeallikas – selleks, et tuvastada anomaaliad andmetes ning hinnata andmete kvaliteeti ja sobivust andmeaida sees;
- andmeaidas – selleks, et veenduda faktis, et allikast tulev andmete kvaliteet on tagatud ka andmeaida siseselt.

Lisaks aitab andmete profileerimine tuvastada, registreerida ja hinnata ettevõtte metaandmeid. Metaandmed on andmed andmeelementide kohta [1]. Metaandmeid on põhjalikumalt kirjeldatud peatükis „Metaandmed“

Andmete profileerimisel kasutatakse kirjeldavat statistikat, arvutatakse näiteks miinimumi, maksimumi, keskmist, moodi ja standardhälvet. Lisaks kasutatakse teisi agregaate nagu summa ja loendus. Metaandmetena kogutavaks infoks võivad olla näiteks andmetüübid, välja pikkused, diskreetsed väärtused, ridade või väljade

unikaalsus, puuduvate väärtuste esinemine või tüüpilised sõnemustrid. Selliseid metaandmeid saab kasutada probleemidel tuvastamisel, näiteks mitte-lubatud väärtused, puuduvad väärtused, duplikaadid, kirjavead ja muutuvate väärtuste esinemine.

Üldiselt on andmete profileerimiseks kolm meetodit [3]:

- veeru profileerimine – analüüsitakse väärtusi veerus, et tuvastada probleeme metaandmetega või sisu kvaliteediga;
- üle tabelite profileerimine – analüüsitakse andmeid üle mitme tabeli, et tuvastada duplikaate või andmete kattuvust ning analüüsida välisvõtmete ja primaarvõtmete seoseid;
- ärireeglite valideerimine – kinnitab andmete vastavust eelnevalt defineeritud reeglitele.

Andmete profileerimise tulemusena saadud informatsiooni ja statistikat võib kasutada [3]:

- kindlaks tegemiseks, kas hetkel olemasolevad andmeid saab kasutada nõutaval eesmärgil;
- pakkumaks andmekvaliteedi mõõdikuid, mis väljendavad andmete vastavust kindlatele standarditele või mustritele;
- hindamaks riski andmete integreerimisel uude asukohta;
- hindamaks metaandmete korrektsust ja vastavust andmetele andmeallikas;
- mõistmaks andmetega seotud probleeme andmeallikates projekti algstaadiumis, et välistada hilisemaid üllatusi;
- pakkumaks ettevõttele ülevaadet ja informatsiooni projekti andmete kohta, mida saab kasutada tuumaandmete haldamiseks. Tuumaandmeteks nimetatakse organisatsioonis hoitavaid andmeid, mis on sõltumatud ja ettevõttele ta tegevuse sooritamisel põhjapanevad alusandmed [8].

3 Andmekvaliteedi probleemid andmete profileerimise etapis

Andmeaidanduse erinevates etappides võib ilmneda erinevaid andmekvaliteedi probleeme. Andmekvaliteedi probleeme on enda uurimustöös „Kirjeldav klassifikatsioon andmekvaliteedi probleemide põhjustest andmeaidas“ kirjeldanud Ranjit Singh ja Dr. Kawaljeet Singth. Eeltoodud uurimustöö autorid on probleemid klassifitseerinud vastavalt erinevatele andmeaidanduse etappidele [7]:

- andmed allikates;
- andmete integratsioon ja andmete profileerimine;
- andmete teisendamine ning töötlemine ETL protsesside abil;
- andmeaida modelleerimine ja skeemide disain.

Antud töös keskendutakse andmekvaliteedi probleemidele andmete profileerimise etapis. Kirjeldatud probleemid pärinevad eelmainitud uurimustööst. Järgnevates alampeatükkides toob autor välja andmekvaliteedi 12 probleemi grupeerituna nende tekkepõhjuste järgi. Iga probleemi järel on autor välja toonud analüüsi teel selgunud probleemi võimaliku lahenduse või ennetusviisi.

3.1 Manuaalne andmete profileerimine

Paljud ettevõtted profileerivad andmeid manuaalselt. Manuaalne lähenemine on praktiline juhul, kui profileerida on vaja vähe veerge ja minimaalselt andmeid. Samuti tagab manuaalne profileerimine parema ettevõttesisese kontrolli ärireeglite ja teenustaseme lepingute üle ning puuduvad kulud, mis kaasnevad uue tehnoloogia õppimisega [9].

3.1.1 Manuaalse profileerimisega seotud probleemid

Enamikul ettevõtetel on tuhandeid veerge ja miljoneid või isegi miljardeid ridu andmeid, mille puhul manuaalne andmete profileerimine nõuaks liiga palju inimeste sekkumist, sisaldaks vigu ja oleks subjektiivne. Manuaalne profileerimine eeldab, et

SQL-päringud koostab inimene, kes oskab väga hästi SQL programmeerimist ning lisaks sellele tunneb väga hästi profileeritavaid andmed ja oskab võimalikke probleeme otsida. Selline lahendus vähendab tõenäosust leida probleeme, mida profileerimise teostaja ei oska ette näha [10].

Põhilised manuaalse profileerimisega seotud probleemid tulenevad inimlikust eksimusest või ebapiisavatest oskustest ja teadmisest. Kasutaja loodud SQL-päringud andmete profileerimise eesmärgil ei tuvasta andmekvaliteedi probleeme. Lisaks on manuaalne andmete profileerimine ajakulukas ja suure tõenäosusega ka puudulik [7].

3.1.2 Võimalik lahendus manuaalse profileerimisega seotud probleemidele

Manuaalse andmete profileerimisega seotud probleemide vältimiseks on võimalik kasutada automaatseid profileerimise tööriistu. Kõige efektiivsemad tööriistad suudavad automatiseerida profileerimise kolm põhilist etappi [10]:

- andmete esialgne profileerimine ja hindamine;
- andmete profileerimise integreerimine automaatsetesse protsessidesse;
- andmete profileerimise tulemuste edastamine andmekvaliteedi ja andmete integratsiooni protsessidele.

Andmete profileerimise tulemused on aluseks andmekvaliteedile ja andmete integreerimise initsiatiivile. Parim lahendus on otsida andmete profileerimise tarkvara, mis võimaldab andmeid parandada, valideerida ja verifitseerida otse profileerimise aruannetes. See võimaldab kombineerida andmete kontrolli ja parandamise faase, mis muudab üldise andmehaldusprotsessi sujuvamaks [10].

Automaatsete profileerimise tööriistade tugevusteks on [9]:

- võimalus profileerida korraga suuri andmehulkasid;
- väiksem inimressursi vajadus andmete haldamiseks ja korrashoiuks;
- tõhusus uute nõuete seadistamise ja käivitamisel;
- võimalus korrastada andmeid reaalajas või planeerida rutiinseid andmekorrastusi;

- võimalus profileerida andmeid nii allikates, andmeaidas kui ka andmevakkades.

Seega tasuks autori hinnangul eelistada automaatseid profileerimise tööriistu, mis vähendavad manuaalse töö vajalikkust ning seeläbi ka inimlike eksimuste tekkimist. Automaatsete profileerimise tööriistade kasutamine võimaldab täpsemat, täielikumat, korratavat ja produktiivsemat profileerimist [3].

Samas tuleks automaatse profileerimise tööriistade puhul arvestada ka võimalike takistustega seoses uue tehnoloogia integreerimisega töökeskkonda, töötamisega kolmandate isikutega, kes ei tunne ettevõtet ja selle vajadusi ning uue tehnoloogia maksumusega [9]. Automaatsete profileerimise tööriistadega seotud probleemidest ning lahendustest räägib autor detailsemalt peatükis „Automaatsed profileerimise tööriistad“

3.2 Automaatsed profileerimise tööriistad

Tänapäeval on saadaval palju erinevaid automaatseid profileerimise tööriistu. Tuntumad neist on Informatica Data Quality, Talend Open Studio for Data Quality, Datamartist ja Toad Data Point. Siiski kõik olemasolevad tööriistad ei tööta samamoodi ning tuleks valida just enda ettevõttele sobiv.

3.2.1 Automaatse profileerimise tööriistaga seotud probleem

Peamine automaatse profileerimise tööriistaga seotud probleem on ebasobiva tööriista valik [3]. Vale tööriista valiku tulemusena võib juhtuda, et tööriista kasutamisel ei saavutata soovitud või vajalikud eesmärgid ning investeering ei tasu ennast ära.

3.2.2 Võimalik lahendus automaatse profileerimise tööriistaga seotud probleemile

Nõuded automaatsele andmekvaliteedi tööriistale peaks välja selgitama hindamise käigus. Hindamise protsess koosneb kolmest osast:

1. Ärivajaduste analüüs
2. Tehniliste vajaduste analüüs
3. Sobiva tarkvara testimine

Järgnevalt kirjeldatakse lähemalt protsessi erinevaid osi.

Ärivajaduste analüüs

Ärivajaduste analüüsi eesmärgiks on välja selgitada milline automaatse profileerimise tööriist sobib ettevõttele kõige paremini. Analüüs koosneb järgnevatest osadest [11]:

- identifitseeritakse äriprotsessid, mida on mõjutanud andmekvaliteet;
- identifitseeritakse andmed, mille kvaliteet on kriitiline nende äriprotsesside edukale teostamisele;
- hinnatakse erinevaid vigu ja andmete puudujääke, mis võivad tekkida;
- kvantifitseeritakse nende vigadega mõju äriotsuste tegemisele;
- prioriseeritakse probleeme vastavalt nende mõjule äriotsuste tegemisel;
- kaalutakse, milliseid andmekvaliteedi parandusi saaks teha, et leevendada mõju äriotsuste tegemisel;

Tehniliste vajaduste analüüs

Tehnilise analüüsi peamine eesmärk on hinnata tarkvara sobivust ettevõtte süsteemi arhitektuuriga. Sobivuse peamised näitajad on [3]:

- kasutatavus – tarkvara on lihtsasti kasutatav kõikidele osapooltele ilma pideva tehnilise nõustamisteta;
- koostöövõime – tarkvara genereerib lihtsasti loetavad raporteid erinevates formaatides, et võimaldada kiiret tulemuste edastamist vajalikele osapooltele;
- otseühendus allikatega – tarkvara on suuteline ühenduma otse allikaga, selmet luua koopiad allika andmetest, mis on aeganõudev ega võimalda analüütikutele ligipääsu kõige värskematele andmetele;
- reeglite genereerimine andmete puhastamiseks – tarkvara võimaldab lisaks probleemsete andmete tuvastamisele ka leitud andmete parandamist etteantud reeglite põhjal või vajaliku info edastamist andmete puhastamisega tegelevale tarkvarale;

- ühilduvus kolmanda osapoole aplikatsioonidega – tarkvara ühildub ETL protsessi ning andmete integratsiooni tarkvaraga, et automatiseerida võimalikult suur osa rutiinsest andmete integratsiooni või migreerimise protsessist;

- laialdane funktsionaalsus – tarkvara pakub vajalikke funktsionaalsusi andmete struktuuri analüüsiks, veergude väärtuste statistika genereerimiseks ja tabelite vaheliste sõltuvuste vastendamiseks.

Sobiva tarkvara testimine

Peale äriprotsesside ning tehnilise keskkonna vajaduste analüüsimist, on võimalik välja selgitada maksimaalselt kolm ettevõtet, kelle tarkvara sobib kõige paremini püstitatud nõuete täitmiseks. Seejärel tuleks testida välja valitud tarkvara soovitud andmeaida keskkonnas. Olles kindlaks määranud võrdluseks oleva andmete kogumiku, saab võrrelda mitte ainult testitava toote jõudlust vaid ka kasutusmugavust ning vastuvõtlikkust oma töötajate seas [11].

Töö autori hinnangul eeltoodud hindamisprotsessi osade korrektsel läbimisel on võimalik leida parim tööriist ettevõtte soovide ja vajaduste täitmiseks.

3.3 Profileerimise ulatus

Andmete profileerimise eesmärk on tuletada informatsiooni andmete kohta. Andmete profileerimine mitte ei otsi ebakorrektseid andmeid, vaid otsib andmeid, mis ei vasta etteantud reeglitele [12]. Andmete profileerimise võib aga olla ebapiisav. Ebapiisav profileerimine väljendub näiteks mõne tabelis olulise veeru profileerimata jätmises, mis omakorda mõjutab antud tabeli põhjal tehtava analüüsi korrektsust ja seeläbi ka andmekvaliteedi hindamist.

3.3.1 Profileerimise ulatusega seotud probleemid

Ebapiisava profileerimise peamine probleem on puudulik veergude, tabelite ning tabeliteülene profileerimine. Lisaks andmete profileerimisele andmeaidas tuleb parima ülevaate saamiseks profileerida andmeid ka andmeallikates, kus on samamoodi oht andmete ebapiisavaks profileerimiseks. Probleemiks on ka manuaalse profileerimise tulemusena saadud informatsioon andmete kohta, mis on suure tõenäosusega ebatäpne ning seeläbi mõjutab andmekvaliteeti [7].

3.3.2 Võimalik lahendus profileerimise ulatusega seotud probleemidele

Ebapiisava profileerimise ennetamiseks tuleb teha põhjalik andmete profileerimine kõikidele tabelitele ja nende veergudele. Järgnevalt toob töö autor välja, milliseid elemente ja kuidas profileerida, et vältida ebapiisavat profileerimist.

Veergude profileerimine

Veergude profileerimine väljendub iga tabeli atribuudi kohta statistika kogumises. Erinevatel veergudel võivad olla erinevad andmetüübid, aga kogutav statistika on üldjoontes sarnane. Kõikide andmetüüpide puhul saab koguda järgmist statistikat [3]:

- puuduvad väärtused ja nende protsentuaalne osakaal;
- olemasolevate väärtuste arv ja nende protsentuaalne osakaal;
- mood – kõige sagedamini esinev väärtus;
- mustrite arv – erinevate täheldatud mustrite arv, näiteks kuupäeva andmetüübiga veergude puhul DD-MM-YYYY või YYYY-MM-DD;
- veeru väärtuste andmetüüp;
- andmete pikkus veerus;
- unikaalsus.

Numbriliste väärtustega veergude puhul saab lisaks eelnevale statistikale koguda ka arvulist statistikat [3]:

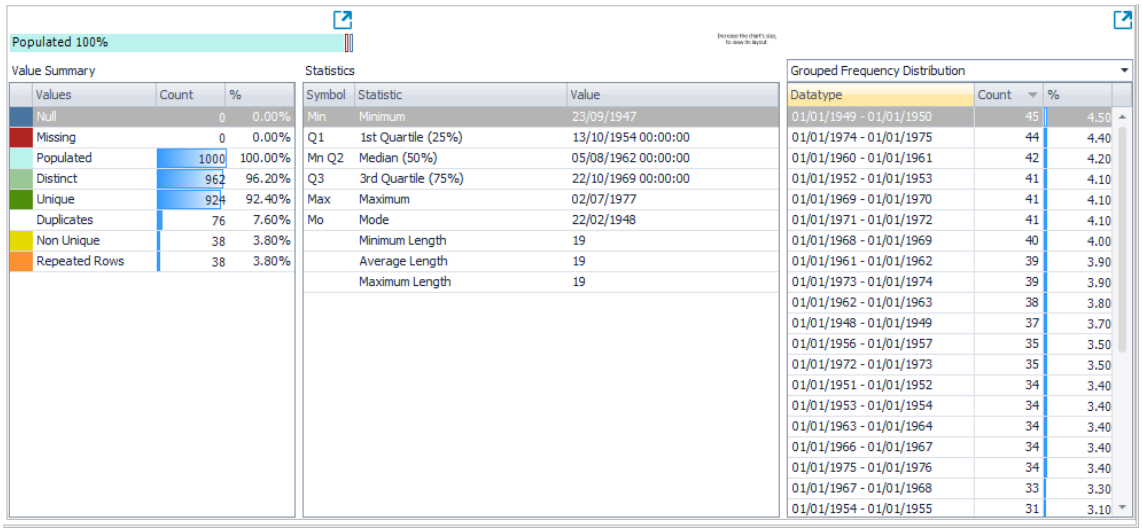
- aritmeetiline keskmine;
- mediaan;
- täpsus;
- standardhälve;
- minimaalne ja maksimaalne väärtus.

Et paremini mõista, millist informatsiooni on võimalik andmete profileerimise tulemusena saada, proovis töö autor automaatset andmete profileerimise tööriista Toad Data Point 3.8. Antud tööriist osutus valituks, kuna sisaldab andmebaasi näiteandmetega, mis võimaldab peale tööriista allalaadimist seda koheselt ka katsetada. Lisaks pakutakse ka tasuta 30 päeva pikkust prooviperioodi. Näidisandmebaasis on tabel nimega *Contacts*, mis sisaldab erinevaid andmeid klientide kohta – nende eesnime, perekonnanime, sugu, telefoninumbrit, töö telefoninumbrit, sünniaega ning e-maili. Tabelis on kokku 100 kirjet ning Joonis 3 on toodud tabeli veerud ning väike osa tabelis sisalduvatest andmetest.

CONTACT...	CUSTOMER_ID	ADDRESS_ID	FIRST_NAME	LAST_NAME	SEX	HOME_PHONE	BUSINESS_PHONE	BIRTH_DATE	EMAIL_ADDRESS
2	17	173	Aubrey	Nesbit	{null}	(445) 491-2452	(234) 951-9688	21/04/1971 00:00:00	Aubrey.Nesbit@Haley.com
3	101	340	Ned	Medlock	F	(557) 983-8069	(301) 655-6589	26/04/1972 00:00:00	Ned.Medlock@F..com
5	69	109	Madeleine	Hatcher	M	(352) 666-3419	(577) 292-1840	27/04/1951 00:00:00	Madeleine.Hatcher@Cleco.com
9	40	309	Echo	Rowls	{null}	(385) 511-6459	(380) 808-1720	09/08/1955 00:00:00	Echo.Rowls@Churchill.com
10	4	366	Xanthe	Braun	M	(332) 498-5891	(606) 536-2061	21/05/1971 00:00:00	Xanthe.Braun@Far-ben.com
13	20	357	Atalaya	Conard	M	(378) 728-5718	(666) 462-3889	01/03/1966 00:00:00	Atalaya.Conard@Alm..com
15	109	172	Theodorick	Beiers	F	(299) 254-2279	(233) 685-7984	17/02/1975 00:00:00	Theodorick.Beiers@Innovus.com
17	55	389	Violet	Smartt	M	(257) 457-5326	(651) 607-8423	06/08/1961 00:00:00	Violet.Smartt@Flachglas.com
18	44	312	Reyburn	Hobson	F	(605) 718-2614	(620) 411-5460	09/03/1948 00:00:00	Reyburn.Hobson@Mentor.com
20	64	331	Vance	Bakker	{null}	(375) 891-7839	(627) 537-4638	24/11/1952 00:00:00	Vance.Bakker@Boston.com
25	70	398	Westley	Lein	{null}	(462) 213-1817	(282) 817-8180	13/03/1962 00:00:00	Westley.Lein@Finaxa.com
27	14	142	Maddox	Devereaux	M	(567) 958-8256	(374) 692-7729	14/05/1948 00:00:00	Maddox.Devereaux@Damart.com
30	107	169	Fillmore	Sorrel	{null}	(787) 490-3732	(651) 323-7332	14/11/1947 00:00:00	Fillmore.Sorrel@Universal.com
33	95	396	Wesley	Day	F	(518) 585-6317	(420) 927-6177	09/08/1965 00:00:00	Wesley.Day@Castle.com
38	116	201	Aileen	Chang	F	(282) 514-2475	(587) 789-9652	11/08/1949 00:00:00	Aileen.Chang@Parque.com

Joonis 3. Ekraanipilt Toad Data Point 3.8 programmist. Näide tabelis *Contacts* sisalduvatest andmetest.

Profileerimise tulemusena saadi erinevat statistikat kõikide veergude kohta. Statistika veeru *BIRTH_DATE* kohta on toodud Joonis 4. Ekraanipilt Toad Data Point 3.8 programmist. Profileerimise tulemusena veeru *BIRTH_DATE* kohta saadud statistika.



Joonis 4. Ekraanipilt Toad Data Point 3.8 programmist. Profileerimise tulemusena veeru *BIRTH_DATE* kohta saadud statistika.

Saadud statistika põhjal saab teha analüüsi, et välja selgitada andmete kvaliteet ning valideerida või genereerida erinevaid ärireegleid. Analüüsist räägitakse täpsemalt peatükis „Profileerimise tulemuste analüüs“.

Lisaks eraldiseisvate veergude analüüsile tuleks profileerida andmeid ka veergudeüleselt. Üle veergude profileerimisel on kolm peamist aspekti [13]:

- võtmete analüüs – vaadeldakse atribuutide väärtusi iga kirje puhul, et teha kindlaks primaarvõtmete kandidaadid;
- sõltuvused – funktsionaalsete sõltuvuste analüüs võimaldab leida varjatud seoseid või varjatud struktuuri andmehulkade puhul. Lisaks võimaldab sõltuvuste analüüs identifitseerida üleliigseid andmeid ja vastandatud väärtuseid ning aitab soovitada võimalusi andmete standardiseerimiseks;
- mustrite, sageduste ja domeeni analüüs – mustrite analüüsiga saab kontrollida, kas andmed on korrektselt formaaditud. Sageduste analüüs aitab autentida andmeallika andmeid. Domeeni analüüs võimaldab valideerida jaotuste väärtusi kindlate andmeelementide vastu. Sellised analüüsid on kasulikud ärireeglite kontrollimiseks ja genereerimiseks.

Üle tabelite profileerimine

Üle tabelite profileerimine annab ülevaate, kuidas erinevate veergude väärtused erinevates tabelites potentsiaalselt ristuvad ja kattuvad. Üle tabelite profileerimise peamised aspektid on [13]:

- välisvõtmete analüüs – aitab tuvastada võtmete suhteid tabelite vahel;
- andmetervikluse kontroll – aitab tuvastada välisvõtmete rikkumisi, mille puhul tütartabelis eksisteeriv kirje ei eksisteeri ematabelis;
- semantiliste ja süntaktiliste erinevuste tuvastamine – näiteks erinevate nimedega veerud sisaldavad samasugust infot või samasuguste nimedega veerud sisaldavad erinevat infot;
- seoste tüüpide analüüs – selgitab välja 1-1, 1-M, 1-0, M-1, M-M ja 0-1 seosed veergude vahel.

Profileerimise tulemusena saadud info põhjal saab teha järeldusi erinevate veergude sobivuse kohta välisvõtmeteks, sõltuvalt profileeritud andmekogude suurusest ja andmekogude kattuvuse astmest [13]. Lisaks võimaldab üle tabelite profileerimine tuvastada kattuvusi veergudevaheliste väärtuse puhul ja võimalikke andmeliiasusi tabelite vahel [3].

Töö autor leiab, et eeltoodud juhiste järgimisel tehtud profileerimine on piisav, et tuvastada kõik olulised probleemid andmetega. Põhjaliku profileerimise tulemusena saadud info põhjal on võimalik genereerida olulised ärireeglid, mille vastu saab hakata andmeid kontrollima.

3.4 Probleemide dokumenteerimine

Probleem on iga stsenaarium, mis mõjutab teenuse või süsteemi usaldusväärtust või kättesaadavust. Probleemid võivad ilmneda erinevatest allikatest ja mõjutada paljusid protsesse. Seega on oluline, et kõik profileerimisel ilmnevad probleemid dokumenteeritakse korrektselt, et võimaldada nende lahendamine ning seeläbi taastada teenuse või süsteemi ootuspärane töötamine [14].

3.4.1 Probleemide dokumenteerimisega seotud probleem

Peamine dokumenteerimisega seotud probleem on puudulik andmete seoste identifitseerimine. Lisaks põhjustavad dokumenteerimata muudatused profileerimise etapis andmekvaliteedi probleeme [7].

3.4.2 Probleemide dokumenteerimine – võimalikud lahendused

Olenemata sellest, kas ettevõttes kasutatakse andmete profileerimiseks automaatset tööriista või profileeritakse andmeid manuaalselt, tuleks profileerimise käigus leitud andmekvaliteedi ja andmete vaheliste seoste probleemid dokumenteerida. Enamik automaatseid profileerimise tööriistu pakuvad tarkvarasiseselt probleemide automaatset dokumenteerimist. Juhul, kui tarkvara seda ei võimalda või profileeritakse andmeid manuaalselt, tuleks välja töötada ühtne probleemide raporteerimise ja dokumenteerimise süsteem. Järgnevalt kirjeldab töö autor, kuidas dokumenteerida erinevaid probleeme, et võimaldada kõigil osapooltel olla kursis hetkel teadaolevate andmekvaliteedi probleemidega ning võimalusel aidata probleemi juurpõhjuse leidmisega.

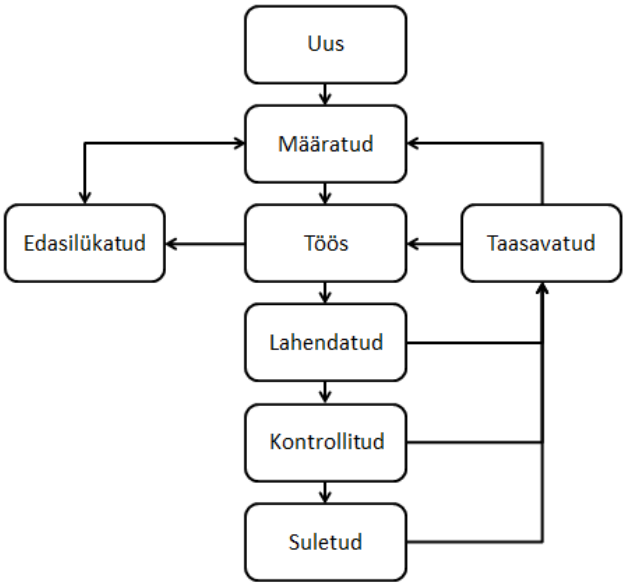
Dokumenteerimise puhul kõige olulisemaks osaks on probleemi kirjeldus, mis sisaldab täpset infot probleemi kohta: kuidas probleem ilmnes, kuidas probleemi taasesitada, millist tulemust oodati ja milline tulemus tegelikult oli. Probleemi taasesitamise lisamine aitab probleemi lahendajal probleemi paremini mõista ja sellele lahendust leida. Kindlasti tuleb kasuks ka võimalus lisada probleemi juurde erinevad lisasid, näiteks ekraanipilte, logifaile või muid olulisi faile. Lisaks tuleb märkida ka probleemiga seotud andmeaida osad, et mõista, milline teenus või lõppkasutaja võib antud probleemist negatiivselt mõjutatud olla.

Probleemide puhul on olulisel kohal ka probleemi tõsidus. Selle määramiseks tuleb välja töötada ühtne süsteem, näide on toodud Tabel 1. Näide võimalikust probleemide tõsiduse määramisest.

Tabel 1. Näide võimalikust probleemide tõsiduse määramisest [3].

Tõsiduse_Id	Tõsidus	Kirjeldus
1	Kriitiline	Süsteem lakkab töötamast, võimalik andmekadu või korrumpeerunud andmebaas, nõuab kohest parandust
2	Kõrge	Suur osa süsteemist on antud probleemi tõttu kasutamatu, tuleks parandada esimesel võimalusel
3	Keskmine	Komponendi või protsessi ebakorrektne funktsionaalsus, võimalik lihtne ajutine lahendus probleemile
4	Väike	Kosmeetilised vead, parandada siis kui aega on

Lisaks probleemi tõsidusele, tuleb välja töötada ka ühtne süsteem probleemi staatuse määramiseks. Üks näide võimalikest probleemide staatustest ning nende määramise järjekorrast on toodud Joonis 5.



Joonis 5. Näide võimalikust probleemide staatuse määramistest [3].

Peale dokumentatsiooni koostamist tuleb kindlaks teha, et antud dokument jõuaks ka õigete inimesteni. Selle lihtsustamiseks võib kasutada erinevaid veajälgimissüsteeme (näiteks Bugzilla või Jira) või edastada dokument e-maili teel. Veajälgimissüsteemide plussiks on reaalajas kommentaaride lisamine, mis võimaldab kergemini saada värskeimat infot probleemi staatuse ja probleemi sisu kohta. Lisaks saavad probleemi uuendamisel kõik probleemiga seotud isikud e-maili teel teavituse, mis annab märku tehtud muudatustest. Dokumentide e-maili teel edastamise puhul tuleks probleemi lahendamisega seotud isikutel igapäevaselt vaadata üle probleemi staatus ja progress

ning edastada vajalik informatsioon seotud inimestele [3]. Üks näide Jira keskkonnas raporteeritud probleemist on toodud lisas 1.

Töö autor leiab, et parim lahendus probleemide dokumenteerimiseks on töötada välja ühtne tõsiduse ja staatuse määramise süsteem ning rakendada see kasutades mõnda veajälgimissüsteemi. See tagab suurima tõenäosusega, et vajalik informatsioon jõuab õigete inimesteni ja õigel ajal.

3.5 Andmete standardiseerimine

Andmete standardiseerimine on protsess, mille järgi erinevates formaatides sisuliselt sarnased andmed muudetakse tavaformaati, mis parendab võrdluse protsessi [15]. Näitena standardiseerimata andmetest võib tuua kuupäevad. Mõned kuupäeva sisestamise formaadid:

- 20.04.2016;
- 20-04-2016;
- 20. aprill 2016;
- 20. aprill ’16;
- 2016-04-20;
- 2016.04.20.

Sisuliselt on tegemist ühe ja sama kuupäevaga, kuid andmeaida süsteem käsitleb neid kui erinevaid väärtusi. Erinevate väärtuste tekkimise peamine põhjus on see, et andmeallikasse tulevad andmed läbi erinevate kanalite, näiteks läbi veebivormide, manuaalse sisestamise või failide laadimise [16].

Andmete standardiseerimise eesmärk on andmete muutmine terviklikumaks ning selgemaks. Terviklikkus kindlustab, et väljundandmed on usaldusväärsed ja seotud andmeid saab identifitseerida kasutades ühist terminoloogiat ja formaati. Selgus väljendub selles, et andmed on kergesti mõistetavad ka nende inimeste poolt, kes ei tegele andmete haldamise protsessiga [17].

3.5.1 Andmete standardiseerimisega seotud probleem

Peamine andmete standardiseerimisega seotud probleem on väär kirjete ja väljade vormingu parsimine ja standardiseerimine tavaformaati. Selle tulemusena on andmekvaliteet negatiivselt mõjutatud, sest mõlema protsessi käigus muudetakse algselt sisestatud väärtust ning kui parsimise või standardiseerimise aluseks võetud reeglid on valed, siis on ka lõpptulemusena saadavad andmed valed. Seega on oluline, et andmeid standardiseeritakse ja parsitakse ainult vajalikke andmeid kontrollitud reeglite järgi.

3.5.2 Võimalik lahendus andmete standardiseerimisega seotud probleemile

Järgnevalt kirjeldab töö autor kuidas lihtsustada andmete standardiseerimisega seotud protsessi. Standardiseerimise eesmärk on reguleerida kriitilised tegevused, mis mõjutavad olulisel määral lõpptulemust. Standardiseerimise protsess on keeruline, kuid samuti standardiseeritav. Toimiva standardini jõudmiseks võiks jälgida järgmisi samme [18]:

1. Probleemi defineerimine – enne probleemi lahendamist tuleb konkreetselt defineerida lahendatav probleem.

2. Meeskonna kokkupanek – probleemi lahendamise juures on abiks sobiv meeskond, mis koosneb inimestest, kes puutuvad kokku probleemile eelnevate ja järgnevate protsessidega. See võimaldab andmeallikaga töötavatel inimestel öelda, milliseid andmeid on võimalik pakkuda ning äriteabe poolel töötavad inimesed saavad esitada nõuded – millises formaadis ja milliseid andmeid vaja läheb.

3. Juurpõhjuste analüüs – lahendada tuleb probleemide juurpõhjuseid ning standardiseerida juhised, mis aitavad tegevust koordineerida ja töötajaid harida.

4. Rakendamine – standardi rakendamise puhul tuleb üle vaadata, kas standard ka reaalselt töötab nii nagu peab. Tuleb koguda tagasisidet töötajatelt ning vastavalt vajadusele ka standardit korrigeerida.

5. Probleemi lahendus – peale standardi rakendamist tuleb kontrollida, kas lahendamise eesmärgiks seatud probleem reaalselt lahenes.

Lisaks standardiseerimisele on oluline ka andmete parsimine. Andmete parsimine on üks andmete korrastamise osa, mille abil jagatakse andmeväljal olevad andmed

väiksemateks osades kindlate reeglite järgi selleks, et andmed oleksid kergemini mõistetavad ja hallatavad. Parsimise eesmärk andmeaidas on üles otsida ja identifitseerida üksikud andmeelemendid andmeallikates ning isoleerida need elemendid sihtkohas. Andmeaidas toimub parsimine ETL protsesside abil.

Üheks keerulisemaks näiteks parsimise kohta on aadress. Aadressi kirjutamisel on erinevaid võimalusi:

- Ehitajate tee 45, korter 6, Tallinn
- Ehitajate tee 45-6, Tallinn, 12612
- Ehitajate tee 45-6, Tallinn, Harjumaa
- Ehitajate tee 45-6

ETL protsess peaks sisaldama algoritme, mis võimaldab erineval kujul olevatest aadressidest välja lugeda vajalikku infot. Eelnevate näidete põhjal peaks ETL protsess suutma saada järgmise info tabelisse *Address*:

Tabel 2. Tabel *Address* peale andmete parsimist

Tänav	Maja_Number	Korteri_Number	Linn	Maakond	Postiindeks
Ehitajate tee	45	6	Tallinn		
Ehitajate tee	45	6	Tallinn		12612
Ehitajate tee	45	6	Tallinn	Harjumaa	
Ehitajate tee	45	6			

Kuna andmete parsimine on üks andmete puhastamise võimalus, tuleb ETL protsessil peale andmete parsimist ka andmed puhastada. Selle jaoks kasutatakse ärireegleid. Antud näite puhul võiks üks reeglitest olla, et kui veeru *City* väärtus on ’Tallinn’, siis ETL protsess sisestab veergu *County* väärtuse ’Harjumaa’ ning lisaks veergude *Street, House_Number* ja *City* väärtuste kombinatsiooni puhul lisab veergu *Postal_Code* väärtuse ’12612’. Nende eelnevalt määratud reeglite täitmisel mõistab ETL protsess, et eeltoodud 4 erinevat näidet on sama aadress erinevates formaatides ning lisab tabelisse *Address* ainult ühe veeru:

Tabel 3. Tabel *Address* peale andmete puhastamist

Tänav	Maja_Number	Korteri_Number	Linn	Maakond	Postiindeks
Ehitajate tee	45	6	Tallinn	Harjumaa	12612

3.6 Metaandmed

Andmeaida metaandmed on informatsioon andmeaidas olevate andmete kohta, mida hoiustatakse ühes või rohkemas spetsiaalses metaandmete hoidlas. Metaandmed sisaldavad [19]:

- informatsiooni andmeaida sisu, asukoha ja struktuuri kohta;
- informatsiooni protsesside kohta, mis varustavad andmeaita uute, ajakohaste, semantiliselt ja struktuuriliselt kooskõlastatud andmetega;
- informatsiooni andmete semantika kohta koos muu informatsiooniga, mis võimaldab lõpp-kasutajal kasutada andmeaidast saadavat informatsiooni;
- informatsiooni andmeaida komponentide ja andmeallikate infrastruktuuri ja füüsiliste omaduste kohta;
- informatsiooni andmeaida turvalisuse, autentimise ja kasutamise kohta, mis aitab andmeaida administraatoril andmeaita vastavalt vajadusele seadistada.

Metaandmeid on profileerimisel nii sisendiks kui väljundiks. Andmete profileerimise protsess parandab ja täiendab protsessi alguseks vajalikke metaandmeid. Seega on sellised metaandmed kõige täielikumad ja täpsemad [12].

3.6.1 Metaandmetega seotud probleemid

Peamiseks probleemiks metaandmete puhul on ebausaldusväärsed ja puudulikud meta-andmed, mis põhjustavad andmekvaliteedi probleeme. Lisaks põhjustab metaandmetega probleeme ka suutmatus hinnata andmete struktuuri, väärtusi ja seoseid enne andmete integratsiooni [7].

3.6.2 Võimalik lahendus metaandmetega seotud probleemidele

Metaandmete kogumine paneb aluse heale andmekvaliteedile. Täpne andmete defineerimine on vajalik selleks, et kontrollida andmete õigsust. Metaandmed iseloomustavad andmeid esitades dokumentatsiooni selliselt, et andmed on kergemini mõistetavad ja lihtsamalt kasutatavad ettevõtte poolt. Metaandmed vastavad andmetega seotud küsimustele: kes, millal, kus, miks ja kuidas [20].

Andmete profileerimise juures aitavad metaandmed paigutada andmed õigetesse kategooriatesse, et teha kindlaks, millised nõuded nendele andmetele kehtivad. Nõuete all võib mõista näiteks informatsiooni veeru andmetüübi, veergu sisestatavate andmete maksimaalse pikkuse, veergude unikaalsuse või veerus puudulike väärtuste lubamise kohta [10]. Mõnede andmete kohta kehtivad mitmed nõuded ja mõnede andmete kohta nõuded üldsegi puuduvad. Ilma korralike metaandmete definitsioonideta on keeruline tagada andmete vastavust nõuetele [20].

Andmete profileerimise tööriistad kontrollivad andmete vastavust nõuetele just metaandmete abil. Reaalsete andmete ja metaandmete erinevus põhjustab kaugeleulatuvaid mõjutusi andmehalduse saavutustes. Seega on oluline, et metaandmed oleks ajakohased ja vastaksid tegelikkusele [10].

3.7 Profileerimise tulemuste analüüs

Andmete profileerimise tulemuseks on hulk informatsiooni ja statistikat andmeallikas olevate andmete kohta, näiteks kui palju erinevaid väärtusi tabeli veerud sisaldavad ja millised need väärtused on. Nende andmete põhjal tuleb teha analüüs ja anda hinnang profileeritud andmete kvaliteedi kohta ning ärireeglitele vastavuse kohta.

3.7.1 Profileerimise tulemuste analüüsimisega seotud probleemid

Profileerimise tulemuste analüüsiga seotud probleemid seisnevad peamiselt tulemuste ebatäpses või pealiskaudses analüüsis. Analüüsi puhul võib puudulikuks jääda saadud kirjeldav statistika, näiteks ridade arv, summa, mood, minimaalsed ja maksimaalsed väärtused, keskmine ja standardne hälve. Lisaks võib olla ebatäpne ka väärtuste vahemike ja jaotuste või läbe analüüs nõutud väljadele. Probleeme tekitab ka suutmatus hinnata andmete struktuuri, väärtusi ja seoseid enne andmete integratsiooni [3].

Probleemide hulga suurusest näha, et profileerimise etapis enim probleeme valmistab just profileerimise tulemusena saadud informatsiooni analüüs.

3.7.2 Võimalik lahendus profileerimise tulemuste analüüsiga seotud probleemile

Profileerimise tulemuste põhjaliku analüüsi puhul keskendutakse peamiselt erinevate andmereeglite kontrollile või loomisele. Andmereeglid on üks alamhulk ärireeglitest, mis defineerivad suhteid erinevate veergude või ridade vahel, mis peavad alati oleme tõesed. Selliste reeglite rikkumine tähendab, et andmed on vigased või ärireegleid, millel andmed põhinevad, reaalselt ei jälgita. Esimesel juhul on põhjuseks valesti sisestatud andmed. Teisel juhul aga on andmed õigesti sisestatud, kuid andmeid ei töödeldud ettevõtte ärireeglite järgi [3]. Mõned näited erinevatest andmereeglitest:

- töötaja palk peab olema vahemikus 500 kuni 10 000 eurot;
- klient peab alkoholi ostmiseks olema vähemalt 18 aastat vana.

Profileerimise tulemusel saadud info põhjal tuleb analüütikul teha põhjalik analüüs. Analüütiku ülesandeks on teostada järgmised etapid [3]:

- analüüsida individuaalseid väärtuseid, et selgitada, kas tegu on antud veergu sobivate väärtustega;
- analüüsida kõiki veerus olevaid väärtusi koos, et leida probleeme unikaalsusega, järjestikkusega ja väärtuste ootamatute sagedustega;
- analüüsida struktuuri reegleid, millega reguleeritakse funktsionaalseid sõltuvusi, primaarvõtmeid, välisvõtmeid, sünonüüme ja duplikaatseid veerge;
- valideerida andmete reegleid, mis kehtivad kindlate andmeridade puhul;
- valideerida andmete reegleid, mis kehtivad kõikide andmeridade puhul ühes äriobjektis;
- valideerida andmete reegleid, mis kehtivad üle mitmete äriobjektide;
- valideerida andmete reegleid, mis kehtivad erinevate seotud äriobjektide puhul;
- valideerida ETL-i ja andmete transformeerimise reegleid.

Peale tulemuste analüüsimist tuleb lisaks olemasolevate andmereeglite kontrollile hinnata ka uute andmereeglite loomise vajalikkust. Andmereeglite loomise puhul tuleks toetuda peamiselt profileerimise tulemustele, sest see tagab, et loodavad reeglid vastavad tegelikele andmetele. Analüütik peaks andmereeglite kontrollimise ja loomise puhul kindlasti tegema koostööd ka andmete lõppkasutajatega, et välja selgitada, millised andmereeglid on neile vajalikud.

Töö autor leiab, et lähtudes eeltoodud analüüsi etappidest võimaldab selline analüüs vältida profileerimise tulemuste ebakorrektset tõlgendamist.

4 Kokkuvõte

Käesoleva töö eesmärgiks oli analüüsida ja leida lahendusi või ennetada erinevad andmekvaliteedi probleeme andmete integreerimise ja profileerimise faasis. Käesolevas töös leiti lahendused probleemidele manuaalse andmete profileerimisega, automaatsete profileerimise tööriistadega, profileerimise ulatusega, dokumenteerimisega, andmete standardiseerimisega, metaandmete ja profileerimise tulemuste analüüsiga. Lahendamata jäid probleemid, mis on seotud vastuoluliste äripotseside hindamisega, andmemustrite analüüsiga, andmete võrdlemisega vastu väliseid võrldusandmeid ning andmete analüüsiga andmeallikates. Kuigi töömahu suuruse tõttu ei jõutud analüüsida kõiki probleeme, leiab töö autor, et tähtsamad probleemid on siiski analüüsitud ning sobivad lahendused leitud.

Töö autor on veendunud, et käesoleva töö tulemusi saab kasutada erinevate suuremate ja väiksemate projektide jaoks, et välja selgitada ja parendada andmete kvaliteeti andmeaidas.

Andmete profileerimist plaanitakse kasutusele võtta ka töö autori ettevõttes ning antud töö tulemusi kasutatakse profileerimise projekti plaanimiseks ning võimalike tekkivate probleemide ennetamiseks. Projektis plaanitakse kasutada leitud lahendusi ühe kindla valdkonna andmete analüüsiks.

Kirjanduse loetelu

[1] H. Vallaste, „e-Teatmik: IT ja sidetehnika seletav sõnaraamat,“ [Võrgumaterjal]. http://www.vallaste.ee/. [Kasutatud 14. aprill 2016].

[2] Wikipedia, „Bill Inmon,“ [Võrgumaterjal]. https://en.wikipedia.org/wiki/Bill_Inmon. [Kasutatud 4. mai 2016].

[3] D. Vucevic ja W. Yaddow, Testing the Data Warehouse Practicum, Trafford Publishing, 2012.

[4] Oracle, „Data Warehousing Concepts,“ [Võrgumaterjal]. https://docs.oracle.com/cd/E11882_01/server.112/e25554/concept.htm#DWHSG001. [Kasutatud 12. aprill 2016].

[5] A. Bashmakova, „Why is ETL testing so important?,“ 30. oktoober 2014. [Võrgumaterjal]. http://www.coherentsolutions.com/blog/why-is-etl-testing-so-important/. [Kasutatud 12. aprill 2016].

[6] M. Rouse, „What is business intelligence?,“ TechTarget, oktoober 2014. [Võrgumaterjal]. http://searchdatamanagement.techtarget.com/definition/business-intelligence. [Kasutatud 14. aprill 2016].

[7] R. Singh ja K. Dr. Singh, „A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing,“ 2. mai 2010. [Võrgumaterjal]. http://www.ijcsi.org/papers/7-3-2-41-50.pdf. [Kasutatud 17. aprill 2016].

[8] Cybernetica AS, „Andmekaitse ja infoturbe leksikon,“ [Võrgumaterjal]. http://akit.cyber.ee/. [Kasutatud 12. mai 2016].

[9] S. Zhao, „Manual vs. automated data validation,“ Experian Data Quality, 9 veebruar 2016. [Võrgumaterjal]. https://www.edq.com/blog/manual-vs.-automated-data-validation/. [Kasutatud 25. aprill 2016].

[10] B. Dorr ja P. Herbert, „Data Profiling: Designing the Blueprint for Improved Data Quality,“ [Võrgumaterjal]. www2.sas.com/proceedings/sugi30/102-30.pdf. [Kasutatud 24. aprill 2016].

[11] D. Loshin, „Buyer's Guide: Choosing data quality tools and software,“ TechTarget, 14 detsember 2010. [Võrgumaterjal]. http://searchdatamanagement.techtarget.com/news/2240025847/Buyers-Guide-Choosing-data-quality-tools-and-software. [Kasutatud 20. aprill 2016].

[12] J. E. Olson, Data Quality: The Accuracy Dimension, San Fransisco: Morgan Kaufmann Publishers, 2003.

[13] SAS, „The Practitioner’s Guide to Data Profiling,“ 2012. [Võrgumaterjal]. http://resources.idgenterprise.com/original/AST-0087746_PractitionersGuideToData.pdf. [Kasutatud 2. mai 2016].

[14] Microsoft, „Process 1: Document the Problem,“ 25. aprill 2008. [Võrgumaterjal]. https://technet.microsoft.com/en-us/library/cc543258.aspx. [Kasutatud 5. mai 2016].

[15] IBM, „IBM Knowledge Center,“ 23 oktoober 2014. [Võrgumaterjal]. https://www.ibm.com/support/knowledgecenter/SSWSR9_11.4.0/com.ibm.mdshs.initiateglossary.doc/topics/r_glossary_standardized_data.html. [Kasutatud 21. aprill 2016].

[16] A. Nelson, „CRM Rehab: How to Standardize Your Data,“ [Võrgumaterjal]. https://www.ringlead.com/resources/ebooks/crm-rehab-how-to-standardize-your-data/. [Kasutatud 21. aprill 2016].

[17] Oracle, „Standardize Data,“ [Võrgumaterjal]. https://docs.oracle.com/cd/E35636_01/doc.11116/e29134/stan_data.htm. [Kasutatud 2. mai 2016].

[18] K. Krinal, „Terviklik kvaliteedijuhtumine: Standardiseerimine,“ [Võrgumaterjal]. http://www.kvaliteedijuhtimine.eu/standardiseerimine/. [Kasutatud 23. aprill 2016].

[19] P. Vassiliadis, „Data Warehouse Metadata,“ [Võrgumaterjal]. http://www.cs.uoi.gr/~pvassil/publications/2009_DB_encyclopedia/DW_metadata.pdf. [Kasutatud 24. aprill 2016].

[20] C. S. Mullins, „Data and Technology Today: Regulatory Compliance and the Importance of Metadata Management, Data Quality, and Data Governance,“ 26. september 2013. [Võrgumaterjal]. https://datatechnologytoday.wordpress.com/2013/09/26/regulatory-compliance-and-the-importance-of-metadata-management-data-quality-and-data-governance/. [Kasutatud 24. aprill 2016].

Lisa 1 – Jira keskkonnas raporteeritud probleem