TALLINNA TEHNIKAÜLIKOOL

Infotehnoloogia teaduskond

Mark Edvard Oliver Oja 178910

Tele2 müügiandmete andmetabelite kvaliteedikontroll ja dokumenteerimine

Bakalaureusetöö

Juhendaja: Kristina Murtazin

Magistrikraad

Tallinn 2022

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: Mark Edvard Oliver Oja

18.05.2022

Annotatsioon

Käesolev töö käsitleb telekommunikatsooniettevõtte Tele2 Eesti filiaali andmejärve müügiprogrammi andmekvaliteedi kontrolli, parendusvõimalusi ja schema andmetabelite dokumenteerimist.

Käesoleva töö eesmärgiks on dokumenteerida 4sale schema Tele2 andmejärves ning teostada andmekvaliteedi kontroll schema olevatel tabelitel, kaardistada võimalikud vead ja pakkuda lahendusi.

Ettevõtte peamiseks eeliseks on kontrollitud ja kvaliteetsete andmete kättesaadavus, mis võimaldab andmekasutajatel teha täpsemaid andmeanalüüse ning ettevõttel on võimalik teha vastavaid järeldusi. Analüüsimeeskond kulutab vähem aega vigade selgitamisele ja võimaldab rohkem keskenduda tulevastele ülesannetele.

Sissejuhatuses välja toodud probleemide lahendamiseks viidi läbi schema aktiivses kasutuses olevate tabelite andmekvaliteedi kontroll.

Täiendavaks andmete kvaliteedikontrolliks kasutas töö kirjutaja SQL-päringut, mis tooks esile võimalikud vigased väljad.

Kasutades TDQM raamistikku töötas töö autor 4sale schema tabelid, et hinnata andmekvaliteedi. Raamistiku tulem edastati vastavatele üksustele, kes omavad avastatud probleemide lahenduse võimekust.

Töö tulemusena valmis Tele2 4sale schema kehtiv dokumentatsioon, mis vastab analüütikute nõuetele nii tabeli sisu kui ka tabelite omavaheliste seoste osas. Lõputöö praktiliseks väärtuseks on lisas 2 oleva andmetabeli dokumentatsioon, mida saab kasutada Tele2 igapäevatöös. Lisaks viidi TDQM raamistikus läbi andmekvaliteedi kontroll 4sale schema aktiivses kasutuses olevatele tabelitele, mille tulemusena edastati kvaliteedipuuduste kõrvaldamiseks pakutud lahendused Tele2 vastavatele osakondadele.

Lõputöö on eestikeelne ja sisaldab 31 lehekülge, 5 peatükki, 7 joonist, 3 tabelit.

Abstract

Data Quality and documentation of EE Tele2 sales data tables.

The present work deals with the data quality control, improvement possibilities and documentation of the data table schema of the data lake sales program of the Estonian branch of the telecommunications company Tele2.

The purpose of this work is to document the 4sale schema in Tele2's data lake and to perform data quality control on the tables in the schema, to map possible errors and to offer solutions.

The main benefit of the company is the availability of controlled and high-quality data, which allows data users to make more accurate data analyzes and the company can draw appropriate conclusions. The analysis team spends less time clarifying errors and allows more focus on future tasks.

To solve the problems outlined in the introduction, data quality checks were performed on the tables in active use in the scheme.

For additional data quality control, the job writer used an SQL query that would highlight possible bad fields.

Using the TDQM framework, the author worked through the data tables data quality. The results of the framework were communicated to the relevant units, which have the capacity to solve the problems identified.

As a result of the work, valid documentation of Tele2 4sale schema was completed, which meets the knowledge requirements of analysts, both regarding the content of the table and the relationships between the tables. The practical value of the thesis is the documentation of the data table i.e. Appendix 2, which can be used in Tele2's daily work. In addition, in the TDQM framework, data quality control was performed on 4sale active tables of the scheme, as a result of which the solutions proposed to eliminate the quality deficiencies were forwarded to the respective Tele2 departments.

The thesis is in Estonian and contains 31 pages, 5 chapters, 7 figures, 4 tables.

Mõisted

Data lake – suure hulga andmete hoidla, eestikeelses tõlkes andmejärv.

Data swamp – andmejärv, mille andmekvaliteet on piisavalt madal, et selles sisalduvaid andmeid ei tohiks kasutada, eestikeelses tõlkes andme soo.

Jira – ülesannete haldustarkvara.

MySQL – andmebaaside haldussüsteem.

Query engine – päringute jooksutamismootor.

Schema – mingi kindla süsteemi andmetabelite kogum.

Story point – ülesande töömahukuse hinnang.

Sisukord

Jooniste loetelu

Tabelite loetelu

1. Sissejuhatus

Käesolev töö käsitleb Telekommunikatsiooni ettevõtte Tele2 Eesti haru (edaspidi Tele2) andmejärve müügiprogrammi 4sale schema andmetabelite dokumenteerimist, andmekvaliteedi kontrolli ja parendamise võimalusi.

1.1. Taust

Kõigist programmidest, mida Tele2 töötajad kasutavad, siirdatakse andmed Exacaster osaühingu (edaspidi Exacaster) poolt hallatavasse andmejärve analüütikutele kasutamiseks. Üks kasutatavatest programmidest on 4sale, mis on Tele2 peamine müügiprogramm, mis on kasutusel kõigis esindustes Eestis. Nimetatud müügiprogramm kuulub ettevõttele Iterato osaühing (edaspidi Iterato), kes on Leedu riigis resideeruv ettevõte, kes haldab seda süsteemi. 4sale on üles ehitatud lepingute põhiselt, mis tähendab, et info mida sealt on võimalik kuvada on seotud mõne kindla müügi- või teenuseosutamise lepinguga.

Tele2 finantsosakonna andmeanalüütikute meeskonna igapäeva töö käigus märgati probleeme 4sale andmete kvaliteediga, millest tulenes vajadus teostada täiendav andmekvaliteedi kontroll, kaardistades peamised puudused.

1.2. Probleem

Probleemiks on äriotsuste mitte õigeaegne vastuvõtmine ja otsuse tulemuste operatiivne hindamise tõhusus. Seda põhjustab ebakorrapärane andmekvaliteet, mille tulemusena tekib ressursside lisakulu, nii töötajate ajalise ressursi kulu kui ettevõttele finantsilised lisakulud. Eriala kirjandusest on teada näiteid ebatäpsetest andmetest põhjustatud lisakuludest. Näiteks on kirjandusest teada panga näide, mis kulutas 100 000 dollarit aastas posti- trüki- ja personalikuludele klientide valede aadresside tõttu. Teise näitena kulutas veebiäri aastas umbes 1 000 000 dollarit postikuludele põhjusel, et andmetes eksisteerisid duplikaadid kliendi nimedes ja klientide aadressid olid vigased. Madal andmekvaliteet põhjustab lisa ajakulu andmete vastavusse viimiseks, süsteemi usaldusväärsuse vähenemist töötajate jaoks, klientide rahulolematust, viivitust süsteemide juurutamisel, lisakulu, saamata jäänud tulu ja vastavusprobleemid. [1], [2]

Tele 2 olukorras avalduvad kolm madalast andmekvaliteedist tulenevat alljärgnevat probleemi:

1. Finantskontrollerid on kaotanud usalduse müügiandmete osas;
2. Analüütikutel kulub lisaaeg andmete omavahelisse vastavusse viimiseks;
3. Klientide rahulolematus on tõusnud, sest nendega võetakse ühendust teemadel, millega neil reaalsuses kokkupuudet ei ole.

Andmeanalüütikud on märganud, et 4sale andmetes, mis jõuavad meie andmejärve, on vigu või puudujääke. Kas on kliendi aadress poolik, näiteks Harjumaa 13 või siis telefoni number lõpetamata, näiteks 372 73. Selliste vigade tõttu läheb analüütikutel rohkesti aega andmete kontrollimiseks ja puuduolevate andmete leidmiseks ja asendamiseks teistelt platvormidelt kogutud andmetega. Tele2 jaoks tähendab see, et analüütik kulutab tööaega välditavale tegevusele ja analüütiku ülesande tellija võib teha järeldusi endiselt valede andmete põhjal.

Olukorra lahendamiseks teostati andmekvaliteedi kontroll 4sale schemas aktiivses kasutuses olevatele tabelitele.

Kuna Tele2 andmejärves oleva dokumenteerimisega hakati tegelema suhteliselt hiljuti, täpsemalt aastast 2020, siis antud töö käigus dokumenteeritakse ka andmejärves olev 4sale schema tabelid. Loodavat dokumentatsiooni saavad edaspidi kasutada analüütikud tulevaste ülesannete lahendamiseks.

Tele2 krediidikontrolli ja finantsprojektide juhi soovist tulenevalt peetakse käesoleval tööl andmekvaliteedi kontrolli fookust põhilepingutel ja seadmemüükidel. Finantskontrollerite probleemiks on ebausaldusväärsed 4sale müügi andmed, kuna need ei ühti laoseisude või ettevõtte poolt üles võetud võlgadega. Puuduliku 4sale dokumentatsiooni tõttu lisatakse loodavatesse raportitesse andmed teistest schemadest, asendades 4sale tabelite puuduvaid andmeid. Selliselt toimides tekib olukord, kus saadakse sarnased andmed teistest keskkondades, mille korral andmete terviklikus ei ole kindel.

1.3. Eesmärk

Kvaliteetne ja kontrollitud andmestik võimaldab ettevõttes teha korrektsemaid andmeanalüüse ja ettevõttel vastavaid järeldusi. Andmekvaliteedi kontrolli peamine ülesanne on tagada eelnimetatud omadused andmestikkudele. Kvaliteetne andmestik võimaldab analüüsimeeskonnal kulutada vähem aega vigade selgitusele ja võimaldab enam keskenduda järgnevatele ülesannetele.

Lähtuvalt eelöeldust on käesoleva töö eesmärgiks dokumenteerida Tele2 andmejärves olev 4sale schema ning selle abil teostada andmekvaliteedi kontroll schemas olevatele tabelitele, võimalike vigade kaardistamiseks ja lahenduste pakkumiseks.

1.4. Töö edasine struktuur

Käesolev lõputöö koosneb järgnevatest osades: metoodika, peamised tulemused, analüüs ja järeldused, kokkuvõte ning lisad.

Objekti kirjeldus, tööriistad, päringu keeled, meetodid ja tööprotsess on välja toodud metoodika osas.

Peamiste tulemuste all on esitatud tehniline dokumentatsioon, mis kirjeldab kuidas töö kirjutaja probleemist tulemusteni jõudis.

Analüüs ja järeldused peatükis selgitatakse probleemi lahendust andmekvaliteedi parandamisel, tehtud töö võimalikud alternatiivid ja tuuakse välja kasu nii ettevõttele kui ka analüütikutele. Samuti on lisatud soovitused, kuidas antud teemaga edasi toimetada.

2. Metoodika

Käesoleva töö eesmärgi täitmiseks kasutati mitmeid andmekvaliteeti kontrolli abistavaid tegevusi ja tööriistu. Valitud tööriistad abistasid andmekvaliteedi kontrolli läbiviimist kui ka andmete dokumenteerimist.

2.1. Objekti kirjeldus

Tele2 andmeanalüütikud kasutavad Jira programmi, et eksisteerivaid ülesandeid hoiustada, töösse võtta ja tekkivaid tulemusi ka tellijale esitada. Töö tulemused esitatakse tavaliselt URL lingina, mis viitab Exacaster lehele tehtud raportile. Andmeanalüütikud hindavad sisse tulnud ülesandeid story pointidega. Story pointid võivad olla 1, 2, 3, 5, 8, 12 ja 20, kus 2 punkti peaks tähendama kuni ühte tööpäeva ja 20 siis kahe nädala töö mahtu (st 80 tundi). Hetkel 4sale'iga seonduvad ülesanded, mida hinnatakse näiteks viie punktiga oleksid teostavad kahega, kui ei peaks läbi viima nii detailset kontrolli ning oleks kasutada kvaliteetne dokumentatsioon.

4sale on müügikeskkond läbi mille tehakse Tele2 kõike lepingutega seonduvat. Andmete osas tähendab see, et 4sale schemas on koos nii kliendi ehk kliendi andmed, sealhulgas email, aadress, nimi, isikukood, kontakt number, kontakteerumise nõusolekud jne kui ka toodete kõik võimalikud andmed, seal hulgas müüdud seadmed, simid, teenuste paketid, lisavarustused ja nende detailsed andmed, näiteks seadme värv, imei (International Mobile Equipment Identity) ehk nii öelda telefoni identifitseeriv kood, hind, mudel, jms.

Kõik need andmed asuvad erinevates andmetabelites, mille vahel eksisteerivad tabelite vahelised seosed. Asjakohase dokumentatsiooni korral oleks võimalik teostada kogu lepingu info väljavõtt tabelitest.

Andmekvaliteedi kontrolli on vaja, et ära hoida *data lake* muutumist *data swampiks*. Sõltumata hetkel tehtavatest andmekvaliteedi pingutustest on teadaolevalt ikkagi tabelites olevad andmed vigased. Kui sellega aktiivselt mitte tegeleda muutub vigaste andmete kogus piisavalt suureks, et *data lake* asemel tuleb hakata kasutama terminit *data swamp* [3]. Mõistagi *data swamp* on analüütikutele väheväärtuslik.

2.2. Tööriistad, keeled

Allolevalt kirjeldatakse kasutatud tööriistad, mis olid vajalikud töö eesmärgi lahendamiseks.

Microsoft Access:. Microsoft Accessi kasutati antud projekti raames 4sale schema diagrammide loomiseks ja nende visuaalina välja toomiseks. Seostest loodud pildid on rakendatavad nii kasutuses olevas dokumentatsioonis kui ka andmete kvaliteedi kontrolli teostamises, kus tabelitest tekitati loogilised plokid, mida oli võimalik üheaegselt kontrollida.

Apache Hadoop: Keskkond mis võimaldab hoiustada andmekogumit, näiteks *data lake* või *data warehouse* (andmeladu). Kõik Tele2 andmed asuvad Exacaster poolt hallatavas Apache hadoop keskkonnas. Apache hadoop on avatud lähtekoodiga raamistik, mida kasutatakse suurte andmekogude hoiustamiseks ja töötlemiseks [4]. Andmejärve kasutatakse paljude eraldiseisvate süsteemide andmete hoiustamiseks, eesmärgiga need kõik koondada ühte kokku. Samuti selleks, et analüütikutel oleks keskkond kus nad saavad ise andmeid pärida ja töödelda ilma ohuta, et see mõjutaks töötavat süsteemi [3].

Apache Impala: Päringumootor. Võimaldab teha päringuid Tele2 andmejärvest kasutades SQL Impala päringukeelt. Impala on Apache Hadoopi avatud lähtekoodiga analüütiline massiivse paralleeltöötluse andmebaas [5], [6]. Kasutades seda päringumootorit, teostati kõik dokumentatsiooni ja andmekvaliteedi kontrollid.

SQL Impala/Hive: Päringukeel. Kõik töö raames andme päringud ja andmetega seonduvad tegevused toimusid just selles keeles, kasutades Apache Impala keskkonda.

phpMyAdmin – Veebipõhine rakendus MySQL üle veebi kasutamiseks. Läbi selle saab hallata andmebaase ja jooksutada MySQL päringuid. Antud rakendust kasutas autor baaside sisemise struktuuri dokumentatsiooni vaatamiseks ja andmebaasi vs andme tabeli sisu kõrvutamiseks.

Excel: Peale, SQL Impala keskkonnas tulemuse saamist oli võimalik Exceliga eksportida tulemused, et neid mugavamalt kui hadoop keskkonnas üle vaadata ja kontrollida.

Confluence: Veebikeskkond, kus Tele2 hoiab kõik võimalikku dokumentatsiooni, mis on kõigile töötajatele kättesaadav. Ka andmeanalüütikud teevad kõik siirdatud tabelite, loodud view'de ja raportide dokumentatsioonid sellesse keskkonda. Confluence on koostöö vikipeedia tööriist, mida kasutavad meeskonnad koostöö tegemiseks ja teadmiste jagamiseks [7]. Käesolevas töös loodud dokumentatsioonid püstitati sellesse keskkonda, kasutamiseks kõigile Tele2 töötajatele. Näide püstitatud dokumentatsioonidele on Lisas 2.

Total Data Quality Management (TDQM) raamistik: Raamistik lähtub andmetest kui tootest, millel on kindel eluiga, algus ja lõpp. Peale parenduste sisseviimist on raamistik ja selle tulemused lõppenud. TDQM-i kohaselt on parendusplaani koostamiseks vajalik olemasolevate probleemide põhjuste analüüsimine. See analüüs on võimalik ainult siis, kui kvaliteedinõuete vaatenurgast saadakse selliste probleemide mõõt, mida on vaja defineerida, mõõta, analüüsida ja parendada [8]. TDQM kasutades on võimalik määrata probleemi kaal ja läbi lisanduvate kontrollide määrata probleemi võimalikud tekkekohad. Autor kasutas seda raamistikku, et teostada andmekvaliteedi kontrolli andme tabelitele.

2.3. Tööprotsessi kirjeldus

Sissejuhatuses välja toodud probleemide lahendamiseks teostati andmekvaliteedi kontroll 4sale schemas aktiivses kasutuses olevatele tabelitele.

Täiendavalt andmekvaliteedi kontrolliks kasutas töö kirjutaja SQL päringut, mis tooks esile võimalikud vigased väljad.

Probleemi lahendus algas 4sale schema aktiivses kasutuses olevate tabelite dokumenteerimisega Confluenci keskkonda. Selleks töötas töö autor läbi aktiivses kasutuses olevad 61 tabelit. Töö dokumentatsiooni protsess algas 4sale schema haldaja Iterato ITAnalyst/Scrum Master konsultatsiooniga.

Aktiivses kasutuses olevad tabelid jagas töö kirjutaja sisu poolest loogilisteks plokkideks, nagu näiteks lepingulisad, põhilepingud, lepingu staatused, toodetega seonduv, jms.

Kasutades TDQM raamistikku töötas töö autor eelnimetatud plokid läbi andmekvaliteedi hindamiseks. Raamistiku tulem edastati vastavatele üksustele, kes omavad esitatud probleemide lahenduse võimekust.

Keerukust lisas ka asjaolu, et andmeanalüüsil oli vaja lisaks andmetabelite omavaheliste seoste infole, hinnata ja mõista alatabelite ülesehituslikke ja struktuurilisi puudujääke, ning leida ja grupeerida sisestusvigu.

3 nädalat (120h) kulus aega andmejärve 4sale schema dokumenteerimiseks, sh. diagrammide loomiseks, ja struktuurist arusaamiseks. Probleemiks võib lugeda asjaolu, et Iteratol puudus ajakohane dokumentatsioon oma andmebaaside kohta. Põhjusel, et autori vajamineva info

päringud Iteratole, Tele2 ja Exacasterile ei andnud täielikku tulemust, taotles autor juurdepääsu otse Iterato andmebaasidele, läbi mille oleks võimalik andmetabelite vahelist seost selgitada. Järgnev nädal kulus Tele2 vajaduste välja selgitamiseks ja vastavalt nendele info leidmiseks. Ettevõtte poolne soov oli fokuseerida tähelepanu põhilepingutele ja teenustele. Nädal kulus erialakirjandusega tutvumiseks andmekvaliteedi teemal. Järgnev 3 nädalat kulus TDQM tsükli teostamisele, sealhulgas analüüsi tulemi vormistamisele, millega ka saavutati lõpuprojekti eesmärk. Põhilepingu seoste ploki kontrollimiseks kirjutati 20 rida SQL koodi, mida 11 tabeli kvaliteedi hindamiseks jooksutati 741 korda, erinevate kontrollide läbi viimiseks. Sarnaselt teostati kontroll ka teenuste ploki osas.

Töö teostamisel lähtus kirjutaja ülikoolis õpitule ja nii ülikoolis kui ka ametialaselt omandatud praktilistele oskustele, saades sellele kinnitust läbitöötatud materjalidest.

3. Peamised tulemused

Peamised tulemused koosnevad kahest osast. Ettevalmistav etapp, mille tulemusel kujunenud dokumentatsioon oli eelduseks põhietapis teostatavale andmekvaliteedi kontrollile.

3.1. Ettevalmistav etapp

Ettevalmistavas etapis teostati 4sale schema dokumenteerimine. Töö eesmärgist tulenevalt oli oluline esmalt dokumenteerida 4sale schema tabelid, et järgnevalt koostada nende põhjal loogilised plokid ja teostada andme kvaliteedi kontrolli.

Töö dokumentatsiooniga algas 4sale ametliku haldaja Iterato töötaja konsultatsiooniga. Mille tulemusel ilmnes vajadus saada juurdepääs Iterato andmebaasidele, et näha sisemist dokumentatsiooni, sh. andmebaaside vahelisi seoseid ja sisemiselt kommenteeritud veerge.

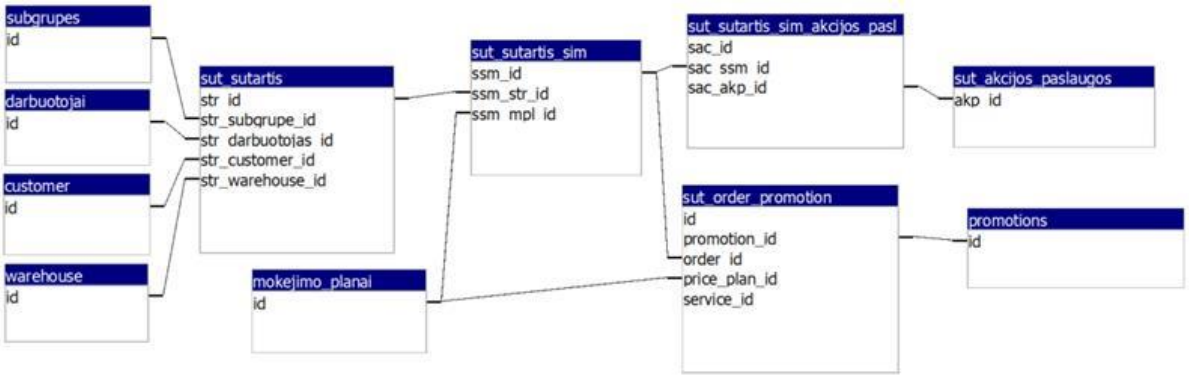
Konsultatsiooni käigus pidid vastuse saama kaks peamist küsimust:

1. Milliseid tabeleid hetkel enam aktiivselt ei täiendata? See tähendab, et millised tabelid on hetkel asendatud uuematega ja jäetud meie andmejärve ajalooliste andmetega tabeliteks. Näiteks viimase suurema 4sale süsteemi uuendusega asendati tabel nimega mjs ja mjs_returns tabelitega contract_phone ja contract_phone_return. Tabelite ülesehitus ja arhitektuur muutus mõne veeru võrra kuid peamine sisu ehk siis seadme müükide ja tagastuste andmed olid samad;
2. Millised seosed eksisteerivad aktiivsete tabelite vahel. Ehk siis vajaminev info, et töö tegija saaks tabelid jagada sisu poolest loogilisteks andme plokkideks.

Konsultatsiooni käigus selgus, et puudub ülevaade tabelitest, mida aktiivselt enam ei täiendata. Töö ülesande lahendamise seisukohalt tähendas see, et autor pidi leidma erinevaid viise kehtiva dokumentatsiooni koostamiseks. Dokumentatsioon koostati olemasoleva osalise dokumentatsiooni, Iterato andmebaaside küljes oleva struktuuri dokumentatsiooni ja andmejärve tabelis oleva info kõrvutamisel 4sale keskkonnas kuvatud lepingute infoga.

Info kõrvutamiseks oli algselt vajadus hinnata käesoleva tabeli sisu, et samasisuline info leida 4sale keskkonnas kuvatud lepingult. Filtreerimise tulemusel, kuvati andmejärves see sisu, mis võimaldaks seostada käesolev tabel juba olemasoleva tabeliga.

Schemas on kokku 90 tabelit, millest 30 ei ole enam aktiivsed, see tähendab, et neile ei lisandu igapäevaselt uusi andmeid ja seetõttu pole neile vaja teostada andmekvaliteedi kontrolli. Allesjäänud 60 täienevad igapäevaselt uute andmetega.



Joonis 1. 4sale schema põhilepingu ploki diagramm.

Dokumentatsiooni loomisel koostas töö autor nii tabelite detailsed dokumentatsioonid kui ka tabelite vaheliste seoste kohta schema diagrammid. Põhjusel, et 90 tabeli vahelisi seoseid näitavat diagrammi on keeruline lugeda, otsustas töö autor jagada tabelid loogilisteks osadeks ja nende kohta luua diagrammid. Näide diagrammist Joonisel 1.

Dokumentatsiooni koostamise ülesehitus jagunes 6 veeru vahel. Neist esimene sisaldab järjekorra numbrit andmejärve tabelis, teine veerg sisaldab andmejärve veeru nimetust tabelis, kolmas sisaldab andmejärve veeru andmetüüpi (string, int, datetime, decimal), neljas veerg sisaldab inglise keelset selgitust andmejärve veerule, viies veerg sisaldab veeru kasutatavust (jah, ei, ei soovita), kuues veerg sisaldab kommentaari andmejärve veeru sisu kohta (näiteks str_pratesimo_tipas 16, mis tähendab Lauatelefoni Numbriliikuvus lepingut). Täpne sõnastus kujunes tööprotsessi käigus ja on esitatud neljandas peatükis (Analüüs ja järeldused).

Veergude sisu valik kujunes autori isiklikul kogemusel töötades Tele2 andmeanalüütikuna.

4sale schema põhilepingu (sut_sutartis) tabeli dokumentatsioon on esitatud Lisas 2.

3.2. Põhietapp

Põhietapis teostati teadaolevate vigade mahu ja päritolu kontroll.

Andmekvaliteedi hindamiseks on loodud erinevaid raamistikke (*frameworke*). Raamistike eesmärk on aru saada andmekvaliteedi hetke tasemest ning välja tuua parendus kohad. [9]

Raamistike alusteooria järgi on andmekvaliteedi hinnangu (alustase) saamiseks soovitatav läbida järgmised tegevused [8]:

- Kliendi registreerimisandmete valmistamisega seotud isikute tuvastamine;
- Olemasoleva andmearhitektuuri kaardistamine;
- Kasutajapiirkondade esitatud andmekvaliteedi probleemide kinnitamine ja teiste tuvastamine;
- Tuvastatud probleemide võimalike põhjuste väljaselgitamine.

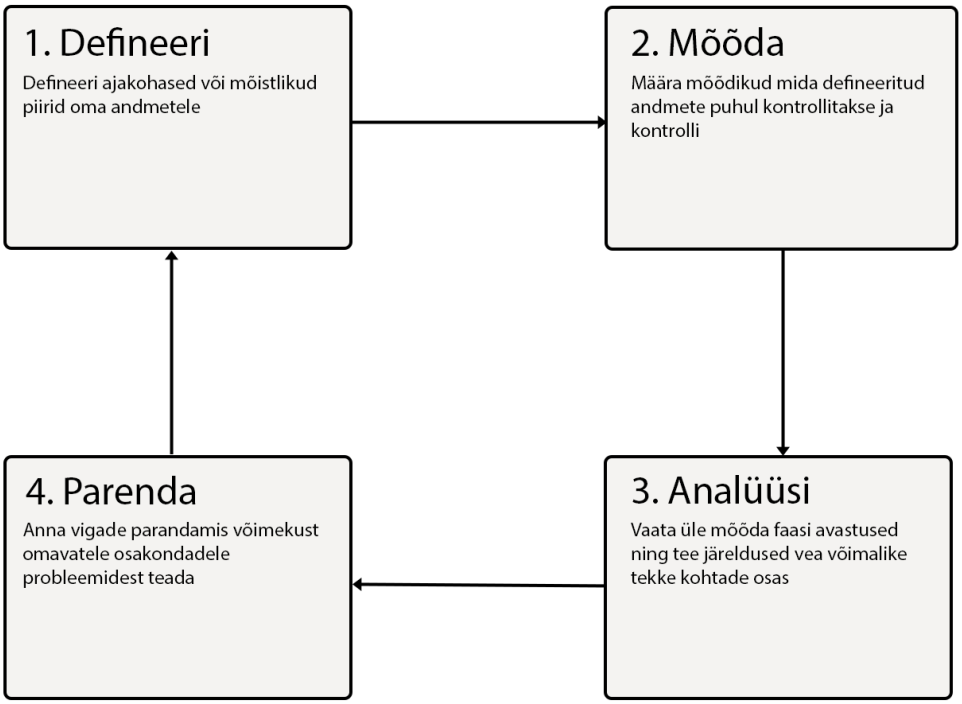
Ettevalmistava etappiga sai tehtud esimesed kaks tegevust, ehk on teada kes tegelevad andmete loomisega ja milline on andmearhitektuur.

Järgmiste tegevuste teostamiseks on vajalik kasutada andmekvaliteedi raamistikku.

Praktikas kasutatavad raamistikud on näiteks:

TDQM (Total Data Quality Management), TIQM (Total Quality Information Management), DWQ (Data Warehouse Quality Methodology), AIMQ (A methodology for information quality assessment), HDQM (Heterogeneous Data Quality Management) [8]

Antud töö käigus kasutati TDQM raamistikku, kuna dokumentatsioon oli eelnevalt koostatud.



Joonis 2. Raamistiku TDQM tsükkel. [9]

Allpool on toodud TDQM raamistiku kirjeldus [8], mis koosneb neljast järjestikusest faasist: Defineeri, Mõõda, Analüüsi ja Parenda (Joonis 2).

Defineeri: Defineeri ajakohased või mõistlikud piirid oma andmetele, millega tegelema hakkad [8]. Kuna raamistik käib kogu andmejärve kohta siis käesolevas töös on definitsiooniks 4sale schema ja andmed peale 2020-12-09, ehk andmed peale viimast suuremat 4sale süsteemi uuendust. Andmete piiritlemine tulenes järgnevast:

1. 4sale schema andmetabelid, sest tegu on peamise müügiandmete koguga, milles oli märgatud vigu. Näiteks kulutati kahe story pointise ülesande peale ligikaudu kaks kuud. Peamised probleemid ajakulu tekkimisel: mittetäielik dokumentatsioon, andmete ebatäpne klassifitseerimine, vajadus ülesande tulemi jaoks pärida vajaminevad lisaandmed teistest schemadest ja kontrollida nende täielikku kattuvust;
2. Andmed alates 09. detsembrist 2020, sest antud kuupäeval toimus viimane suurem 4sale süsteemi uuendus. Mille käigus muutus ka aktiivses kasutuses olevate tabelite koosseis.

Mõõda: Mõõda faasis määratakse mõõdikud mida defineeritud andmete puhul kontrollitakse. Näiteks täpsus, kas kliendil on kirjas korrektne aadress , või täielikkus, kas kliendil on üldsegi kirjas aadress [8]. Antud töö käigus kasutas töö kirjutaja täpsuse ja täielikkuse mõõdikuid defineeritud andmejärve kontrollimiseks. Täpsemalt on mõõdikud kirjeldatud neljandas peatükis (Analüüs ja järeldused).

Analüüsi: Analüüsi faasis vaadatakse üle Mõõda faasi avastused ning tehakse järeldused vea võimalike tekke kohtade osas [8]. Mõõda faasis leitud vigade andmete mahu kaardistamine, vea allikate tuvastamine ja statistika koostamine. Sealhulgas süsteemsete ja mittesüsteemsete vigade eristamine käesoleva töö käigus. Süsteemsete vigade esinemise korral võttis töö kirjutaja kasutusele kolmanda mõõdiku, milleks oli toorandmete võrdlemine siirdetabeli andmetega ehk Usaldusväärsuse mõõdik.

Parenda: Parenda faasis antakse vigade parandamise võimekust omavatele osakondadele probleemidest teada. Lisaks võib tekitada raporti või automaatset teavitussüsteemi, mis toob esile sarnased vead tulevikus [8]. Kuna käesolevas töös on tegu Iterato hallatava programmiga, esitab töö kirjutaja võimalikud lahendused kontaktiks olnud Iterato esindajale.

Esimene plokk, mis läbis TDQM tsükli oli „Uute lepingute“ plokk. See koosneb 11 tabelist ja vaatluse all oli 600 000 rida.

Schema plokkide kontrolli jaoks oli tarvilik uuesti defineerida TDQM tsükkel.

Definitsioon: Vaatluse all on 11 omavahel ühendatavat tabelit: sut_sutartis, sut_sutartis_sim, sut_sutartis_sim_akcijos_paslaugos, sut_akcijos_paslaugos, mokejimo_planai, sut_order_promotion, promotions, warehouse, customer, subgrupes, darbuotojai. Need tabelid kujundavad kokku põhilepingu info. Põhilepingu kooseisu kuuluv info on järgmine: kliendi andmed, osutatavad teenused, sim kaardi info, osutatavate teenuste soodustused, telefoni paketid, millisest laost mõni toode välja läks (näiteks sim kaart ja esinduse ladu), millises poes müük toimus ja kes oli müüja. Lisaks piirangud, et vaatluse all on lepingud alates 2020-12-10 ja lepingud pole loodud TEST Esinduses ega ICIT Support esinduses.

Mõõda: Mõõtmise jaoks kasutas autor dokumentatsioonis kirja pandud tabelite vahelisi seoseid. Sellest tulenevalt loodi baaspäringu kood, mis on joonisel 3.

```
SELECT *
From (SELECT * From sut_sutartis
where str_datalaikas > "2020-12-10"
) sut

LEFT OUTER JOIN sut_sutartis_sim sim on sim.ssm_str_id = sut.str_id
LEFT OUTER JOIN sut_sutartis_sim_akcijos_paslaugos vas on vas.sac_ssm_id = sim.ssm_id
LEFT OUTER JOIN sut_akcijos_paslaugos vas_value on vas_value.akp_id = vas.sac_akp_id
LEFT OUTER JOIN mokejimo_planai pp on pp.id = sim.ssm_mpl_id
LEFT OUTER JOIN sut_order_promotion order_prom on order_prom.price_plan_id = pp.id and order_prom.order_id = ssm_id
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
LEFT OUTER JOIN warehouse warehouse on warehouse.id = sut.str_warehouse_id
LEFT OUTER JOIN customer customer on customer.id = sut.str_customer_id
LEFT OUTER JOIN subgrupes subgrupes on subgrupes.id = sut.str_subgrupe_id
LEFT OUTER JOIN darbuotojai darb on darb.id = sut.str_darbuotojas_id

where subgrupes.centro_pavadinimas not in ( "TEST ESINDUS", "ICIT Support")
;
```

Joonis 3. Baaspäringu kood, põhilepingu andmekvaliteedi kontrolli jaoks.

Üldine kontroll veergudele oli algselt võtta Distinct väärtus, et teada saada, millised väärtused veerus üldse esinevad. Kombineerides seda Where funktsiooni erinevate klauslitega oli võimalik esile tuua potensiaalseid vigu.

```
SELECT DISTINCT str_id
From (SELECT * From sut_sutartis
where str_datalaikas > "2020-12-10"
) sut
LEFT OUTER JOIN sut_sutartis_sim sim on sim.ssm_str_id = sut.str_id
LEFT OUTER JOIN sut_sutartis_sim_akcijos_paslaugos vas on vas.sac_ssm_id = sim.ssm_id
LEFT OUTER JOIN sut_akcijos_paslaugos vas_value on vas_value.akp_id = vas.sac_akp_id
LEFT OUTER JOIN mokejimo_planai pp on pp.id = sim.ssm_mpl_id
LEFT OUTER JOIN sut_order_promotion order_prom on order_prom.price_plan_id = pp.id and order_prom.order_id = ssm_id
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
LEFT OUTER JOIN warehouse warehouse on warehouse.id = sut.str_warehouse_id
LEFT OUTER JOIN customer customer on customer.id = sut.str_customer_id
LEFT OUTER JOIN subgrupes subgrupes on subgrupes.id = sut.str_subgrupe_id
LEFT OUTER JOIN darbuotojai darb on darb.id = sut.str_darbuotojas_id

where length(cast(str_id as string)) < 5

```

Joonis 4. Lepingu numbri kontroll sut_sutartis tabelist.

Joonisel 4 on esitatud lepingu numbri kontroll, mis toob välja, kas eksisteerib lepingu numbrit, mis on lühem kui 5 tähemärki. Sellise kontrolli peamiseks eesmärgiks oli selgitada, kas esineb teistsuguses formaadis või tühje väärtusi. Antud näites tagastas päring No Results ehk Lepingu numbri veerus ei ole ei tühjasid ega „vales“ formaadis lepingu numbreid.

Juhul kui esimeses kontrollis avastati vigu, oli vaja määrata selle vea osakaal. Selleks kasutas autor Count funktsiooni koos Group By funktsiooniga koos erinevate Where klauslitega, millega saaks välistada õigeid ridu, et võimaldada vigaste ridade kokku loendamist.

```
SELECT count(*)
From (SELECT * From sut_sutartis
where str_datalaikas > "2020-12-10"
) sut

LEFT OUTER JOIN sut_sutartis_sim sim on sim.ssm_str_id = sut.str_id
LEFT OUTER JOIN sut_sutartis_sim_akcijos_paslaugos vas on vas.sac_ssm_id = sim.ssm_id
LEFT OUTER JOIN sut_akcijos_paslaugos vas_value on vas_value.akp_id = vas.sac_akp_id
LEFT OUTER JOIN mokejimo_planai pp on pp.id = sim.ssm_mpl_id
LEFT OUTER JOIN sut_order_promotion order_prom on order_prom.price_plan_id = pp.id and order_prom.order_id = ssm_id
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
LEFT OUTER JOIN warehouse warehouse on warehouse.id = sut.str_warehouse_id
LEFT OUTER JOIN customer customer on customer.id = sut.str_customer_id
LEFT OUTER JOIN subgrupes subgrupes on subgrupes.id = sut.str_subgrupe_id
LEFT OUTER JOIN darbuotojai darb on darb.id = sut.str_darbuotojas_id

where subgrupes.centro_pavadinimas not in ( "TEST ESINDUS", "ICIT Support")
and (length(split_part(str_contact_name, " ", 4)) > 1 or length(split_part(str_contact_surname, " ", 4)) > 1)
;
```

Joonis 5. Count ja Where funktsiooni kombinatsiooni päringu kood vea osakaalu kontrolliks.

Äriklientide nime vormistamisel täheldati mitmeid erinevaid vormistusviise. Seetõttu oli vajadus kontrollida erinevate sisestusviiside hulka. Joonisel 5 on välja toodud kliendi nime kontroll. Esitatud kontroll välistas kõik need, kellel ei olnud ees- või perekonnanime lahtris neljakohalist nime (kontroll tulenes ühest ettevõtte nime sisestus tüüpist).

```
SELECT --count(str_email)
str_email, count(str_email)
From (SELECT * From sut_sutartis
where str_datalaikas > "2020-12-10"
) sut
LEFT OUTER JOIN sut_sutartis_sim sim on sim.ssm_str_id = sut.str_id
LEFT OUTER JOIN sut_sutartis_sim_akcijos_paslaugos vas on vas.sac_ssm_id = sim.ssm_id
LEFT OUTER JOIN sut_akcijos_paslaugos vas_value on vas_value.akp_id = vas.sac_akp_id
LEFT OUTER JOIN mokejimo_planai pp on pp.id = sim.ssm_mpl_id
LEFT OUTER JOIN sut_order_promotion order_prom on order_prom.price_plan_id = pp.id and order_prom.order_id = ssm_id
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
LEFT OUTER JOIN warehouse warehouse on warehouse.id = sut.str_warehouse_id
LEFT OUTER JOIN customer customer on customer.id = sut.str_customer_id
LEFT OUTER JOIN subgrupes subgrupes on subgrupes.id = sut.str_subgrupe_id
LEFT OUTER JOIN darbuotojai darb on darb.id = sut.str_darbuotojas_id

--where str_sutarties_sub_tipas = ""
where length(cast(str_email as string)) < 5
GROUP BY str_email

;
SELECT *
From sut_order_promotion order_prom
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
ORDER BY order_prom.id desc
```

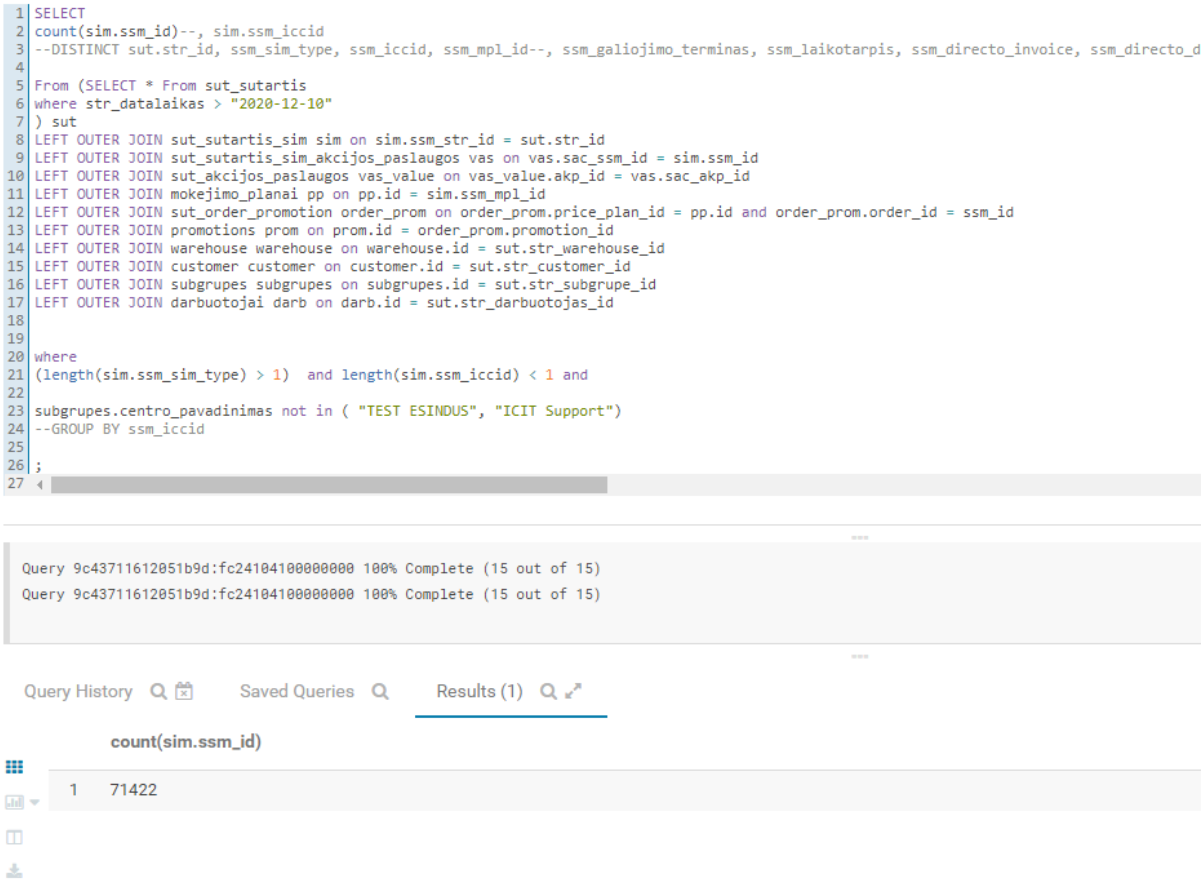
```
Query 3747fb268b79a172:a6c55fa300000000 100% Complete (15 out of 15)
Query 3747fb268b79a172:a6c55fa300000000 100% Complete (15 out of 15)
```

Query History | Saved Queries | Results (2)

	str_email	count(str_email)
1		66476
2	.	1

Joonis 6. Group By ja Where funktsioonide kombinatsiooni päringukood Emaili veeru vigade osakaalu kontrolliks.

Joonisel 6 on esitatud emaili veeru kontrolli tulemus, mille kohaselt on esitatud kõik emaili variatsioonid, mis on lühemad kui 5 tähemärki. Päringu vaste näitab, et 66476 real on email puudu ja ühel on emailiks punkt.



```
SELECT
count(sim.ssm_id)--, sim.ssm_iccid
--DISTINCT sut.str_id, ssm_sim_type, ssm_iccid, ssm_mpl_id--, ssm_galiojimo_terminas, ssm_laikotarpis, ssm_directo_invoice, ssm_directo_d

From (SELECT * From sut_sutartis
where str_datalaikas > "2020-12-10"
) sut
LEFT OUTER JOIN sut_sutartis_sim sim on sim.ssm_str_id = sut.str_id
LEFT OUTER JOIN sut_sutartis_sim_akcijos_paslaugos vas on vas.sac_ssm_id = sim.ssm_id
LEFT OUTER JOIN sut_akcijos_paslaugos vas_value on vas_value.akp_id = vas.sac_akp_id
LEFT OUTER JOIN mokejimo_planai pp on pp.id = sim.ssm_mpl_id
LEFT OUTER JOIN sut_order_promotion order_prom on order_prom.price_plan_id = pp.id and order_prom.order_id = ssm_id
LEFT OUTER JOIN promotions prom on prom.id = order_prom.promotion_id
LEFT OUTER JOIN warehouse warehouse on warehouse.id = sut.str_warehouse_id
LEFT OUTER JOIN customer customer on customer.id = sut.str_customer_id
LEFT OUTER JOIN subgrupes subgrupes on subgrupes.id = sut.str_subgrupe_id
LEFT OUTER JOIN darbuotojai darb on darb.id = sut.str_darbuotojas_id


where
(length(sim.ssm_sim_type) > 1)  and length(sim.ssm_iccid) < 1 and

subgrupes.centro_pavadinimas not in ( "TEST ESINDUS", "ICIT Support")
--GROUP BY ssm_iccid

;
```

```
Query 9c43711612051b9d:fc24104100000000 100% Complete (15 out of 15)
Query 9c43711612051b9d:fc24104100000000 100% Complete (15 out of 15)
```

Query History Saved Queries Results (1)

	count(sim.ssm_id)
1	71422

Joonis 7. Count funktsiooni päringukood sim kaardi info ridade kontrolliks.

Joonisel 7 esitatud klausel toob välja need read, kus eksisteerib info väljastatud sim kaardi tüübi kohta, aga on puudu sim kaardi iccid (Integrated Circuit Card Identifer ehk sim kaardi unikaalne kood selle identifitseerimiseks). Päringu vastus näitab, et sim kaardi iccid on puudu 71422 real.

Analüüsi: Alljärgnevas analüüsis kasutatakse tähenduses „kasutud“ veerge, kus esinevad andmed on kas konstantsed või konstantselt puudu. Tähenduses „probleemsed“ veerge, kus esinevates andmetes eksisteerib vähemalt ühte tüüpi viga. Tähenduses „korras“ veerge, kus andmed on kvaliteetsed.

Tabel sut_sutartis – Põhiline lepingu informatsioon, kliendi andmed, kuupäev ja ühendused tabelitesse sut_sutartis_sim, warehouse, customer, subgrupes, darbuotojai. Antud tabelis on 64 veergu, millest 13 on probleemidega, 27 on kasutud ja 24 on korras.

Tabelis 1 on esitatud põhilepingu andmete enam esinenud vigade avaldumine. Selgub, et kõige rohkem vigu esineb aadressi välja ja kliendi nime sisestamisel.

Tabel 1. sut_sutartis tabeli enam esinenud vigade ülevaade

Veerg	**Probleem**	**Hulk**
Aadressi väli	Aadressi väli märkimata	57 000
	Automaatsüsteemi poolt valideerimata, aadressi vale formaadi tõttu	64 000
Kliendi nimi	Erakliendil märkimata	51 000
	Ettevõtte nimi: ettevõtte tüüp eraldi ees- või perekonnanime lahtris	600
	Ettevõtte nimi: ettevõtte nimi ees- või perekonnanime lahtris koos ettevõtte tüübiga	7291
	Ettevõtte nimi: Ettevõtte nimi koos kontaktisiku nimega ees- või perekonnanime lahtris	3000
Suhtluskeel	Suhtluskeel märkimata	48 000
Kontakt telefon	Kontakt telefon märkimata	47 000

Tabel sut_sutartis_sim – Põhilepingu informatsioon, lepingus osutatavad teenused (paketid, sim kaardid, lisateenused nagu mobiili id, kindlustus, go3, jms) ja ühendus tabelitesse sut_sutartis_sim_akcijos_paslaugos (lisateenuste info), mokejimo_planai (pakettide info) ja sut_order_promotion (antud real oleva teenuse soodustuse info). Antud tabelis on 44 veergu, millest 7 on problemaatilised, 26 on kasutud ja 11 on korras.

Tabelis 2 on esitatud põhilepingu osutatud teenuste andmete enam esinenud vigade avaldumine. Selgub, et kõige rohkem vigu esineb sim kaardi tüübi ja iccid veergudes. sut_sutartis_sim tabeli vigade kontrollist ilmnesid ka tabeli ülesehituslikud vead. Näiteks on duplikaat veerg kontakttelefoni andmete jaoks ning juhul kui lepingus ei ole järelmaksu on järelmaksu lõppkuupäevaks 1970-01-01.

Tabel 2. sut_sutartis_sim tabeli enam esinenud vigade ülevaade.

Veerg	**Probleem**	**Hulk**
Sim kaardi toote kood	Toote kood märkimata	120 000
Sim kaardi iccid	Iccid märkimata	164 000
Kontakttelefon	Kontakt telefon märkimata	88 000
	Kontakt telefonil i täht lõpus	420
Järelmaksu lõpp kuupäev	Lõpp kuupäev enne lepingu sõlmimise algust	31
Järelmaksu periood	Periood 0	136

Tabel sut_sutartis_sim_akcijos_paslaugos – Lisateenuste nimi ja ühendus tabelisse sut_akcijos_paslaugos. Tabelis on 5 veergu ja sisu on korras.

Tabel sut_akcijos_paslaugos – Lisateenuste detailid. Tabelis on 27 veergu, millest 8 on kasutud ja 19 on korras.

Tabel mokejimo_planai – Pakutavate pakettide info. Tabelis on 34 veergu, millest 14 on kasutud, 2 on id'd mis viitavad tabelitele mida ei ole Tele2 andmejärves, 18 on korras.

Tabel sut_order_promotion – Soodustuse ja teenuse vahetabel, läbi selle ühenduvad teenus, leping, pakett ja soodustus. Ühendub edasi promotions tabelisse. Tabelis on 8 veergu, mis on korras.

Tabel promotions – Kõik võimalike soodustuste tüübid. Soodustuse periood ja väärtus. Tabelis on 8 veergu, millest 1 on kasutu ja 7 korras.

Tabel warehouse – Kõik võimalike ladude nimistu, nende nimi ja kood. Tabelis on 3 veergu, mis on korras.

Tabel customer – Klientide tabel, milles on kliendi isikukood või registrinumber ja kas on tegu era- või ärikliendiga. Tabelis on 3 veergu, millest 1 on problemaatiline ja 2 on korras. Probleem on isikukoodi/registrinumbri veeruga, kus on 256k tühja rida.

Tabel subgrupes – Müügi sooritanud poe info. Tabelis on 31 veergu, millest 11 on kasutud, 11 on teadmata infoga, 1 viitab tabelile mis puudub andmejärves ja 8 on korras.

Tabel darbuotojai – Lepingu loonud müüja info. Tabelis on 20 veergu, millest 6 on problemaatilised, 2 on teadmata infoga, 8 on kasutud ja 4 on korras. Lisaks ilmnesid ka tabeli ülesehituslikud vead. Näiteks on duplikaat veerg töötaja nime andmete jaoks

Tabel 3. darbuotojai tabeli enam esinenud vigade kirjeldus.

Veerg	**Probleem**	**Probleemi kirjeldus**
Töötaja nimi	Formaat	Esines peamiselt kolm erinevat formaati: suured tähed, väiksed tähed, enamus suured ja täpitähed väiksed
Allkirja pdf viide	Puuduv sisu	Tuhandel real on töötaja allkirja pdf viite väärtuseks NULL
Töötajate grupeering	Grupeeringu kirjeldus	Töötajate grupeeringute märkimise formaat on kooskõlastamata, näiteks ühel grupil on 3 erinevat sisestus viisi: Sales specialist, müügispetsialist, MÜÜGISPETSIALIST.

Kokkuvõtvalt plokis on 247 veergu, millest 109 on korras, 27 on vigadega, 95 on kasutud, 3 viitavad väljapoole andmejärve ja 13 on endiselt teadmata infoga. Duplikeeritud infot on korras veergudest 7, see tähendab, et 7 veergu kuvavad infot, mis on juba mujal ära märgitud.

Kogutud ja analüüsitud info põhjal saab öelda, et plokis kontrollitud 247 veerust on ligikaudu 40% kasutud veerud.

Parenda: Ilmnenud probleemid nii sisestusel kui ka andmetabelite ülesehituses esitati koos kirjelduse, mahu näitude ja näidistega Tele2 Business development Managerile edaspidiseks analüüsiks probleemide lahendusele. Eesmärgiga vähendada analüütikute ajakulu ja tõsta esitatavate raportite andmekvaliteeti.

Edastatud parendus ettepanekud:

Konkreetse lahendusena käesolevast analüüsist ärikliendi nime formaadi probleemile, saab pakkuda ärikliendi lepingu vormi juurdearenduse loomist 4sale keskkonda.

Aadressvälja formaadi probleemi lahendusena saab välja pakkuda võimalust ühendada aadressi väli eksisteerivate maps arendustega, see ühtlustaks formaati ja eemaldaks edasise valideerimise nõude.

Töötaja nime ja grupeeringu probleemi lahenduseks saab kontrollida, mis keskkondadest keskkondadest info kokku tuleb ja nende andmebaaside sissekirjutus formaat ühtlustada, kokku leppida grupeeringu formaat.

Lisaks ploki kohta käivatele probleemidele avastas käesoleva töö autor ka puudused tabelite ülesehituses. Olulisemad ja praktilist väärtust omavad ettepanekud edastati Iteratole:

1. Edaspidi peaksid olema tabeli veeru nimetused rahvusvahelise praktika kohaselt inglise keelsed;
2. Edasiste tabelite loomisel kontrollida, et info mida kuvatakse ei eksisteeriks juba teistes tabelites ehk, et ei oleks duplikeeritud infot, mis lihtsalt koormab andmejärve;
3. Schema siseselt välistada tabelites duplikeeritud veeru nimesid, näiteks paljud tabelid algavad veeruga id. Probleem sellega on see, et päringukoodis, mis kasutab rohkem kui ühte select funktisiooni tekib selle tõttu error „AnalystException: duplicated inline view column alias“. Selle vältimiseks on mõistlik nimetada veerud, mille üldnimi eksisteerib ka teistes tabelites eesliitega tabelinimetus_veerunimetus;
4. Selgitada mis põhjusel eksisteerivad sisutühjad veerud Tele2 andmejärves. Kui selgub, et nad ei peakski andmejärves olema siis koostöös Exacasteriga ümber teha igapäevased andmesiirdamise loogikad;
5. Ühtlustada numbriline jah – ei kontroll, näiteks customer_type mis indikeerib kas klient on era- või äriklient. Tabelis sut_sutartis tähistab 1 eraklienti ja 2 äriklienti, samas kui tabelis s12_priedai, mis hoiab lepingu lisade informatsiooni, on customer_type 1 eraklient ja 0 äriklient.

Kõik eelnevalt nimetatud probleemid ja võimalikud lahendused on edastatud eelnevalt nimetatud Tele2 Business development Managerile.

4. Analüüs ja järeldused

Käesolevas töös on analüüsitud millised olid eesmärgi saavutamiseks teostavate praktiliste tegevuste alternatiivid, sammuti saab esile tuua ärilist kasu, peamisi tulemusi, tulemuste valideerimist ja tulevasi tegevusi.

4.1. Praktilised tegevused

Käesoleva töö eesmärkide täitmiseks teostati järgnevad peamised tegevused: dokumentatsiooni koostamiseks vajaliku info saamine, ajakohase dokumentatsiooni koostamine, raamistiku valik andmekvaliteedi teostuseks ja andmekvaliteedi raamistiku rakendamine.

All järgnevalt on välja toodud iga tehtud tegevuse põhjendus, positiivsed ja negatiivsed küljed ja võimalikud alternatiivid.

1. Dokumentatsiooni koostamiseks vajaliku info hankimine:

Dokumentatsiooni loomiseks vajalikku informatsiooni päris töö autor algselt Iteratolt. Kuna Iteratolt ei õnnestunud saada infot soovitud ulatuses, võttis autor ühendust Exacaster, Tele2 poolsete Iterato kontakt isikutega, teiste andme analüütikute (ka töölt lahkunud, kuid konsultatsiooni tasemele jäänud analüütikutega) ja viimaks vaadeldes otse Iterato algandmebaase. Hetkel kehtiv dokumentatsioon koostati kasutades kõiki eelnimetatud komponente, sest igas komponendis oli midagi mida teistes ei olnud. Antud tegevusele alternatiivi ei eksisteerinud. Positiivsed küljed on, et nüüdsest eksisteerib kõige ajakohasem ja kõigist võimalikkudest info allikatest ehk komponentidest kokku pandud kõike täielikum dokumentatsioon. Negatiivne on see, et puudub kontrollallikas ja endiselt tundmatuks jäänud veerge ei ole võimalik välja selgitada ilma Iterato arendajate sekkumiseta.

2. Ajakohase dokumentatsiooni koostamine:

Dokumentatsiooni formaadi loomisel lähtus autor nii enda kui ka teiste andme analüütikute kogemusest ja soovidest. Eelnev dokumentatsioon koosnes kolmest veerust: Field, mis tähistas veeru nime, Field_type, mis tähistas veeru andmetüüpi ja Description, kus oli erineva kvaliteediga selgitust. Loodud dokumentatsioon aga koosneb 6 veerust.

Lisati positsiooni number, mis tähistab veeru esinemise järjekorranumbrit andme tabelis, eesmärgiga lihtsustada veeru üles leidmist, kergesti kontrollida tabeli veergude arvu ja märgata mainimata muudatusi tabelite sisus.

Lisati veerg „Useable“, milles on välja toodud kas veeru sisu on korras, probleemidega, tundmatu või kasutu (sisu on konstantne). Veeru eesmärk on anda vastus küsimusele kuigi veerg on tabelis, kas ma saan seda kasutada?“

Lisati kommentaari veerg, milles on välja toodud veeru sisu täpsustus, kui see on vähegi vajalik, näiteks kui on staatuse veerg, on välja toodud staatuse number ja selle tähendus. Eesmärgiga lihtsustada arusaadavust ja vältida teostatavaid kontrolle veergudele, et selgitada numbrilist väärtust tähistamaks mingit kindlat tüüpi tunnuseid, näiteks customer_type 1 = eraklient, 2 = äriklient.

Täiendati Description (selgitus) veergu, lisades sinna tabelite vaheliste ühenduste puhul viide tabelile ja selle veerule. Näide seosest Lisa 2 rida 5 ehk subgrupes_id veerg viitab subgrupes tabeli id veerule.

Alternatiivide valikut kitsendas fakt, et dokumentatsioon peab saama püstitatud Confluence keskkonda. Ainus alternatiivsus mida valida sai oli väljatoodava info kogus dokumentatsioonis. Loodud dokumentatsiooni veergudest jäi välja veeru sisu näide. Selle eesmärk oleks olnud välja tuua sisu formaat. Näiteks kas telefoni number algab 372 algusega või mitte. Lisaks jäi välja veeru kohta käiv formaadi ühtlustamise koodi näide. Veeru sisu oleks indikeerinud, et tabeli veeru sisus eksisteerib formaadi probleeme ja koheselt näidanud võimalikku lahendust. Samuti jäi välja eesti keelne veeru selgitus, mis oleks tõstnud veeru sisu arusaadavust. Nimetatud veerge ei lisatud sooviga vältida dokumentatsiooni „kirjuks“ muutumist.

3. Andmekvaliteedi teostus raamistiku ehk metoodika valik:

Käsitsi teostavate andmekvaliteedi kontrollideks on loodud erinevaid raamistikke, näiteks TDQM (Total Data Quality Management), TIQM (Total Quality Information Management), DWQ (Data Warehouse Quality Methodology), AIMQ (A methodology for information quality assessment), HDQM (Heterogeneous Data Quality Management). [8]

Kaks peamist metoodikat on TDQM ja TIQM, mis on universaalsemad ja millest kaudselt tulenevad ka uuemad ja eelnevalt väljatoodud metoodikad. Need on kaks metoodikat, mida toetab enamik andmekvaliteedi tööriistu, mis hõlbustavad nende rakendamist. Need

raamistikud toetavad kõiki andmekvaliteedi tsükleid, kuid mitte konkreetset faasi, näiteks AIMQ piirdub andmete kvaliteedi hindamise faasiga. Need on üldised, st neid saab rakendada mis tahes keskkonnas ega ole mõeldud konkreetsesse konteksti, näiteks DWQ keskendub andmelao keskkonnale. [8], HDQM on mõeldud just ettevõtetele, mille põhitegevuseks on juhtmevaba pihuarvutite müük ja seotud teenused kliendisoovi registreerimiseks [10]. HDQM raamistiku ülesehitus oleks kaudselt sobinud antud töö lahendusse, kuid sobilikumad raamistikud töö eesmärgi lahendamiseks olid TDQM ja TIQM.

TDQM raamistikust on detailselt räägitud peatükis 3.2.

TIQM koosneb kuuest tsükli osast, mis algab andmetega tutvumisega

TIQM tsükli protsessid [11]:

Protsess P1 – hindab andmete tähendust ja andmestruktuure – see viitab vajadusele teada andmedefinitsiooni, et mõõta andmebaasi info kvaliteeti.

Protsess P2 – hinda infokvaliteeti – analüüsib kvaliteeti, hallates informatsiooni kui vara.

Protsess P3 – mõõta ebakvaliteetsest teabest tulenevaid kulusid – see määrab kindlaks äritegevuse tulemuslikkuse mõõdikud, arvutab seotud teabe ja puuduva teabe kvaliteedi maksumuse, tuvastab kliendisegmendid, arvutab väärtuse kliendi jaoks ja arvutab teabe väärtuse.

Protsess P4 – andmete ümbertöötamine ja puhastamine – see parandab puuduva kvaliteedi sümptomitele mõjuvat infotoodet. Selle protsessi käigus muudetakse defektsed andmed vastuvõetava kvaliteeditasemega andmeteks.

Protsess P5 – teabe genereerimisega seotud protsessikvaliteedi parandamine - kontrollib kvaliteedihindamise etapis leitud probleeme, analüüsib põhjuseid, kavandab ja viib ellu parendusprotsesse, et vältida defekte, tegutseb otseselt probleemi põhjustele. Täiustused võivad hõlmata muudatusi äriprotsessides ja infosüsteemides või andmete täiustamise protsessi enda rakendamist.

Protsess P6 – teabekvaliteedi keskkonna loomine (tegevuskava) – see esindab juhiseid ja kultuurilisi nõudeid, mis on vajalikud teabekvaliteedi pideva parandamise keskkonna säilitamiseks. Seetõttu on see teiste protsesside aluseks ja määratleb DQ rakendamise tegevuskava.

Raamistiku valik: Põhjusel, et TIQM raamistik alustab dokumentatsiooni loomisega, mis käesolevas töös loodi ettevalmistavas etapis. TIQM raamistiku kolmas protsess tegeleb vigadest tuleneva kulu mõõtmisega, mis oleks ületanud käesoleva töö eesmärkide skoopi. Neljas protsess tegeleb andmete puhastamisega, millega tekib vajadus koormata andmejärve uute tabelitega, kus on puhastatud tabelis ainult paar veergu, alusandmete puhastamiseks puudub töökirjutajal juurdepääs. Mainitud probleemide tõttu otsustas töö autor valida TDQM raamistiku.

4. Andmekvaliteedi raamistiku kasutus:

Valitud raamistiku võimalikud mõõtmed on [9]:

Täpsus: õiged, veatud andmed, järjepidevad esitus ja sisu;

Ajatu: andmed on piisavalt ajakohased, et neid saaks kasutada konkreetses kliendiülesandes;

Täielikkus: kasutamiseks piisavalt andmeid;

Usaldusväärsus: mõiste, mis on seotud andmete usaldusväärsuse ja neid salvestava ja töötleva süsteemiga. Samuti, on see seotud viiteallikate usaldusväärsus osas;

Esitus: andmete mõistmise ja tõlgendamise lihtsus, järjekindel ja kokkuvõtlik esitus.

Käesolevas töös leidsid kasutust Täpsuse ja Täielikkuse mõõtmed. Alternatiivsete mõõtmed oleksid olnud Ajatus, Usaldusväärsus ja Esitus. Ajastus ei leidnud kasutust, sest Tele2 finants kontrollerid kasutavad praeguseni kõiki kontrollitud andmeid. Usaldusväärsus ei leidnud kasutust, sest fundamentaalsed probleemid said avastatud dokumentatsiooni koostamise ajal. Esitus ei saanud kasutust, sest andmed said tõlgendatud ja seletatud dokumentatsiooni loomise ajal.

TDQM mõõda tsükliosas otsustas autor kontrollida andmeid plokkidena, eesmärgiga anda tabelitele kontekst, mis tekitaks arusaadava andmete kogumi. Alternatiiv oleks olnud kontrollida tabeleid üksikult, mis oleks segmenteerinud ja osaliselt lihtsustanud mõõda tsükliosa tegevust. See valik oleks raskendanud arusaadavust ja võtnud ära võimaluse teostada mõningaid kontrolle. Näiteks kontroll, mis läbis mitut tabelit ja kindlalt võtnud ära läbiva kontrolli, et vaadeldavad andmed ei oleks olnud loodud test keskkonnas. Otsuse positiivne külg seisnes selles, et vea avastusel oli võimalik teostada 4sale keskkonnas vaatluskontroll kasutades lepingu kättesaamiseks andmeid mis olid juba juurde lisatud. Otsuse negatiivne külg

seisnes selles, et sut_sutartis_sim tabel, mis tõi välja kõik lisateenused eraldi ridadel, tekitas osades tabelites välja toodud andmetele duplikaadid, näiteks tabel subgrupes, mis tõi välja lepingu loonud esinduse. Mainitud negatiivset külge vähendas teadmine, et kontrollitav baas on piisavalt suur, et õigete ja „vigadega“ andmete osakaal peaks ennast ära tasakaalustama.

4.2. Praktiliste tegevuste äriline kasu

Praktiliste tegevuste ärilist kasu saab hinnata mitmel tasandil – nii igapäevaste tööülesannete täitmisel, kui juhtimisotsuste tegemisel. Dokumentatsioon, mille autor koostas on koos diagrammidega leidnud kasutust vähemalt kolmes analüütikutele määratud ülesandes, mille töösse võtmisel toimus nende ülesannete story pointide ümber hindamine. Lisaks lihtsustati päringu koodi kahes rapordis, kus dokumentatsiooni abil sai koodist eemaldada teistest schemadest lisatud andmed.

Peale müügi andmete andmekvaliteedi kontrolli tegi töö autor finants kontrolleritele uue rapordi, mille andmed kattuvad kontrollerite erinevate kontrollallikatega, tõstes sellega kontrollerite otsuste kvaliteeti.

Peale muudatuste sisse viimist tõuseb Tele2 andmejärve 4sale schema kvaliteet eelnevaga seoses kasvab ka analüütikute loodud raportite kasutajate järelduste kvaliteet.

4.3. Peamised tulemused

Käesolevas töös saab esile tuua kolm peamist tulemust:

1. Dokumentatsioon – Täiendati olemas olevat ja koostati uut dokumentatsiooni 4sale schema aktiivses kasutuses olevatele tabelitele ja nende kõik võimalikele ühendus kohtadele;
2. Tabelite kvaliteedi kontroll – Teostati 4sale schema aktiivses kasutuses olevatele tabelitele kvaliteedi kontroll, toodi välja parendus kohad ja täiendati dokumentatsiooni vigade kohta;
3. Schema põhimõttelised probleemid – Avastati ja toodi esile schema tabelite vahelised põhimõttelised probleemid, mida oleks võimalik parenda või vältida tulevaste tabelite lisamisel.

4.4. Tulemuste valideerimine

Valideeritud tulemustest saab töö kirjutamise ajal välja tuua ainult loodud dokumentatsiooni. Kindlus selle kvaliteedi üle tulenes kahest teadmisest:

1. Töö autor lõi tänu dokumentatsioonile uue seadme müügi rapordi, mille sisu kontrollisid finantskontrollerid ja märkisid korrektseks;
2. Täiendatud dokumentatsiooni on töö autor tutvustanud teistele analüütikutele, sellele järgnevalt on dokumentatsioon saanud nende poolt kasutuse võtu ja heakskiidu.

4.5. Tulevased tegevused, kuidas edasi

1. Koostöös Tele2, Iterato ja Exacasteriga läbiarutada ja võimalusel sisse viia vajaminevad muudatused 4sale andmete kvaliteedi tõstmiseks;
2. Laiendada andmekvaliteedi kontrolli teistesse schemadesse;
3. Rakendada 4sale schemale andmekvaliteedi kontrolli teostavat rakendust näiteks Talend *Data integrity and data governance*, Alation *Data Catalog*. Valitud rakendust ühendatakse andmejärvega ja see hakkab teostama iseseisvalt andmekvaliteedi kontrolle. Rakendus kasutab selleks masinõpet, mis võrdleb tabeli siseselt veergudes olevaid andmeid, tuues välja read, mis ei ühti enam esinenud formaadiga. [12], [13]. Talendil on eelis, sest Tele2 juba kasutab Talendi teist moodulit *Application and API integration.*

5. Kokkuvõte

Käesoleva töö käigus teostati Tele2 andmejärve 4sale schema andmekvaliteedi kontrolli ja dokumenteerimist. Olemasolev dokumentatsioon ei osutunud piisavalt efektiivseks, sest esines piisavalt puudusi, et nõuda puuduva informatsiooni uurimist. Ebakvaliteetse andmestiku kasutamine osutus analüütikutele liialt ajamahukaks. Sellest kujunesid töö eesmärkideks selgitada 4sale schema tabelite seosed ning sisu, mille alusel oli võimalik luua kehtiv dokumentatsioon ja teostada aktiivsetele tabelitele andmekvaliteedi kontroll.

TDQM raamistikku kasutades koostas autor andmekvaliteedi kontrolli dokumenteeritud tabelitele. Raamistiku jaoks tehti päringuid nii mySQL kui ka SQL Impala päringukeeltes.

Töö tulemusel valmis Tele2 4sale schema kehtiv dokumentatsioon, mis vastab analüütikute teadmiste nõuetele, nii tabeli sisu kui ka tabelite omavaheliste seoste kohta. Praktilise väärtusena lõputöös saab lugeda, Lisas 2 esitatud andmetabeli dokumentatsiooni, mis saab kasutust Tele2 igapäeva töös. Lisaks teostati TDQM raamistikus andmekvaliteedi kontroll 4sale schema aktiivsetele tabelitele, mille analüüsi tulemusel, sai kvaliteedi puuduste likvideerimiseks välja pakutud lahendused edastatud, vastavatele Tele2 osakondadele.

Kuigi käesolevas töös kasutas autor andmekvaliteedi hindamiseks TDQM raamistikku, saab sellega teada ainult hetke seisu ja muudatuste nõuded. Püsivamaks lahenduseks oleks vaja kasutusele võtta konstantset andmekvaliteedi kontrolli teostava rakenduse.

Töö autori, Tele2 Krediidikontrolli ja finantsprojektide juhi ja Finants kontrollerite hinnangul leiti käesoleva töö tulemus praktilist väärtust omavaks ja firmale kasu toovaks.

Töö tulemusel püstitatud nõuded täideti ja valminud dokumentatsioon on Tele2 töötajatele kättesaadav igapäevaseks kasutamiseks.

Kasutatud kirjandus

[1] T. O'Brien, M. Helfert ja A. Sukumar, „The Value of Good Data- A Quality Perspective A framework and discussion," Proceedings of the 15th International Conference on Enterprise Information Systems, Frace, 2013. [Kasutatud 15 aprill 2022].

[2] L. Sebastian-Coleman, „The Importance of Data Quality Management," %1 Meeting the Challenges of Data Quality Management, 2022. [Kasutatud 15 aprill 2022].

[3] J. Herritz, „4 Key Factors for Data Quality on a Data Lake (OR: How to avoid the data swamp)," 2021. [Võrgumaterjal]. Available: https://www.miosoft.com/solutions/files/4_Key_Factors_for_Data_Quality_on_a_Data_Lake-MIOsoft.pdf. [Kasutatud 22 märts 2022].

[4] AWS Amazon, „What is Hadoop?," [Võrgumaterjal]. Available: https://aws.amazon.com/emr/details/hadoop/what-is-hadoop/. [Kasutatud 1 aprill 2022].

[5] M. Fotache, V. Greavu-Şerban, I. Hrubaru ja A. Tică, „Big Data Technologies on Commodity Workstations: A Basic Setup for Apache Impala," CompSysTech'18: Proceedings of the 19th International Conference on Computer Systems and Technologies, New York, 2018. [Kasutatud 22 aprill 2022].

[6] M. Kornacker, „Impala: A Modern, Open-Source SQL Engine for Hadoop," [Võrgumaterjal]. Available: https://2013.berlinbuzzwords.de/sites/2013.berlinbuzzwords.de/files/slides/Impala%20tech%20talk.pdf. [Kasutatud 16 aprill 2022].

[7] Atlassian, „Confluence: Features & Functions," Atlassian, 14 July 2020. [Võrgumaterjal]. Available: https://confluence.atlassian.com/confeval/confluence-evaluator-resources/confluence-features-functions. [Kasutatud 15 aprill 2022].

[8] M. Francisco, S. N. A. Souza, E. G. L. Campos ja L. Souza, „Total Data Quality Management and Total Information Quality Management Applied to Costumer

Relationship Management,“ 2017. [Võrgumaterjal]. Available: https://www.researchgate.net/publication/321954023_Total_Data_Quality_Management_and_Total_Information_Quality_Management_Applied_to_Costumer_Relationship_Management. [Kasutatud 2 märts 2022].

[9] Data Science Central, „What is Data Lake and How to Improve Data Lake Quality,“ 22 May 2019. [Võrgumaterjal]. Available: https://www.datasciencecentral.com/what-is-data-lake-and-how-to-improve-data-lake-quality/. [Kasutatud 26 veebruar 2022].

[10] C. Batini, D. Barone, F. Cabitza and S. Grega, "A Data Quality Methodology for Heterogeneous Data," International Journal of Database Management Systems, vol. 3, no. 1, pp. 60-79, 2011. [Kasutatud 22 Veebruar 2022].

[11] L. P. English, "The TIQM® Quality System for Total Information Quality Management: Business Excellence through Information Excellence," in MIT Information Quality Industry Symposium, 2009. [Kasutatud 1 märts 2022].

[12] Alation, „Data Catalog,“ Alation, [Võrgumaterjal]. Available: https://www.alation.com/product/data-catalog/. [Kasutatud 13 märts 2022].

[13] Talend, „Data Catalog Solutions,“ [Võrgumaterjal]. Available: https://www.talend.com/products/data-catalog/. [Kasutatud 13 märts 2022].

Lisad

Lisa 1

Mina, Mark Edvard Oliver Oja

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Tele2 müügiandmete andmetabelite kvaliteedikontroll ja dokumenteerimine", mille juhendaja on Kristina Murtazin.

1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.

3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

Lisa 2: sut_sutartis dokumentatsioon

Nr	Column_name	Type	Eng	Useable	Comment
1	str_id	bigint	Agreement id	Yes	OK
2	str_prekybos_vt_nr	string	Warehouse code	Yes	OK
3	str_msisdn	int	One msisdn that this contract applies to	Yes	(dont recommend using, has blank rows)
4	str_eil_nr	bigint	Table row number	Unknown	OK
5	str_subgrupe_id	int	Shop id, subgrupes.id	Yes	OK
6	str_darbuotojas_id	int	Employee id, darbuotojas.id	Yes	OK
7	str_datalaikas	timestamp	Agreement date	Yes	OK
8	str_processing_time	timestamp	Action processing time	No	Useless (1970-01-01)
9	str_sutarties_tipas	int	Customer type	Yes	OK, (1 RES, 2 BUS)
10	str_sutarties_sub_tipas	string	Customer subtype	Not complete	Blanks (i.e. res, diffrent business, t2)
11	str_klt_id	int	Customer type	Yes	OK, (1 RES, 2 BUS)
12	str_adresas	string	Full aadress	Not complete	Blanks, different formats
13	str_manually_adresas	int	If address is manually inserted	Yes	OK
14	str_adresas_saskaitoms	string	Full address for invoices	Not complete	Blanks, different formats
15	str_manually_adresas_saskaitoms	int	If address is manually added	Yes	OK
16	str_adresas_delivery	string	Delivery address	No	Useless, Blank after 2020-12-10
17	str_email	string	Customer Email	Not complete	Blanks mostly OK
18	str_language	string	Prefered language	Not complete	Blanks mostly OK
19	str_contact_surname	string	Contact surname	Not complete	Blanks mostly OK, many BUS formats
20	str_contact_name	string	Contact name	Not complete	Blanks mostly OK, many BUS formats

21	str_kontaktinis_telefonas	string	Customer contact number	Not complete	Blanks, mostly OK
22	str_laikinas_kontaktinis	string	Customer temporary contact number	Not complete	Duplicate of str_kontaktinis_telefonas
23	str_position_number	string	Unknown	No	Useless (Blank)
24	str_department_number	string	Unknown	No	Useless (Blank)
25	str_office_number	string	Unknown	No	Useless (Blank)
26	str_kiti_duomenys	string	Unknown	No	Useless (Blank)
27	str_avansas	string	Unknown	No	Useless (0)
28	str_skelbti_viesai	int	Unknown	No	Useless (0)
29	str_gauti_reklama	int	Unknown	No	Useless (0)
30	str_mnp	int	Ported msisdn count	Yes	OK
31	str_vp	int	Võrgusisene üleminek kõnekaardilt lepingulisele	Yes	OK
32	str_p2p	int	Unkown	No	Useless (0)
33	str_ag_komentarai	string	Unkown	No	Useless (Blanks)
34	str_aktyvavimo_trukme	bigint	Contract activation duration in hours	Yes	OK, töötlemine -> aktiveeritud billingusse
35	str_status	int	Agreement status (connects to doc_status_description)	Yes	OK
36	str_invoice	bigint	Unkown	No	Useless (NULL)
37	str_invoice_deleted	bigint	Unkown	No	Useless (NULL)
38	str_perpildyta_str_id	bigint	shows initial contract id in case of refill	Yes	OK
39	str_descriptive_data	string	Some descriptive data	Yes	OK, from it can be read out, how many msisdn were in the contract
40	str_pildymo_tipas	int	Unknown	No	Useless after 2020-12-10, (0)
41	str_pratesimo_tipas	int	contract subtype	Yes	OK: 0 - sega ostuleping, 2 - ümbervormistuslepin g, 9 - ainult seadme ostuleping, 16 - Lauatelefoni NL, 21 - Era-arve, 26 - Püsiühendus
42	str_account_no	bigint	Billing account nr (0 = juhuklient)	Yes	OK
43	str_cash_or_card	string	Unknown	No	Useless, (Blank)
44	str_charge_taken	string	Unknown	No	Useless, (Blank)

45	str_eshop	int	If contract is from eshop	Yes	OK, (1 yes / 0 no)
46	str_adr_id	bigint	Internal Id, connected to addr table not in our datalake	Not complete	possibly used by siebel ingestion_id
47	str_adr_saskaitoms_id	bigint	Internal Id, connected to addr table not in our datalake	Not complete	possibly used by siebel ingestion_id
48	str_validated_address	int	0-not valid address 1- valid 2-trusted	Yes	OK
49	str_kliento_tipas	string	Unknown	No	Useless (1)
50	str_account_msisdn	int	Unknown	No	Useless (NULL)
51	str_is_kklt	int	Unknown	No	Useless (0)
52	str_e_signature	int	contract e-signing	Yes	OK, (0 - without e-signature, 1 - created, 2 - sent, 3 - received)
53	str_annex_e_signature	int	contract e-signing	Unknown	(0 - without e-signature, 1 - created, 2 - sent, 3 - received, 11 - ?, 12 - ?, 13 - ?)
54	str_dealer_sent	int	Unknown	No	Useless (0)
55	str_fake_ba	int	Unknown	No	Useless (1)
56	str_eshop_contract_nr	int	Unknown	No	Useless (0)
57	str_confirmation_type	int	describes agreement activation flow (0 - default flow, 1 - external system will confirm agreement for processing)	Yes	OK, 1 kui on alustatud mingisugune tarne
58	str_leasing_id	string	Unknown	No	Useless, (Blank)
59	str_customer_id	int	Connection id, customer.id	Yes	OK
60	str_warehouse_id	int	Connection id, warehouse.id	Yes	OK
61	str_by_phone	int	Unknown	No	Useless, (NULL)
62	str_by_phone_number	int	Unknown	No	Useless, (NULL)
63	str_by_phone_date	timestamp	Unknown	No	Useless, (1970-01-01)
64	org_str_id	bigint	Unknown	No	Useless, (NULL)