DATAN LAADUNHALLINNAN KÄYTTÖÖNOTTO ENERGIA-ALAN YRITYKSESSÄ

Diplomityö

Lappeenrannan–Lahden teknillinen yliopisto LUT

Tuotantotalouden diplomityö

2022

Aino Solin

Tarkastajat: Professori Timo Kärri

Tutkijatohtori Lasse Metso

TIIVISTELMÄ

Lappeenrannan–Lahden teknillinen yliopisto LUT

School of Engineering Science

Tuotantotalouden koulutusohjelma

Aino Solin

Datan laadunhallinnan käyttöönotto energia-alan yrityksessä

Diplomityö

2022

76 sivua, 21 kuvaa, 12 taulukkoa ja 1 liite

Tarkastajat: Professori Timo Kärri, Tutkijatohtori Lasse Metso

Avainsanat: Datan laadunhallinta, datan laatu, käyttöönotto

Diplomityössä tutkittiin datan laadunhallinnan käyttöönottoa energia-alan yrityksen Helenin data-alustalla sekä salkkuennusteprosessissa. Tutkimusta edelsi vuonna 2021 toteutettu nykytila-analyysi datan laadusta ja sen hallinnasta, jossa havaittiin puutteita muun muassa salkkuennusteprosessissa hyödynnetyn datan, kuten sopimustietojen, laadussa. Työn tavoitteet ja rajaukset laadittiin nykytila-analyysissa tunnistettujen puutteiden sekä valitun toteutuskumppanin viitekehyksen pohjalta. Tavoitteena oli ottaa datan laadunhallinnan menetelmiä käyttöön data-alustalla sekä kehittää toimivia datan laadunhallinnan yhteistyö- ja toimintamalleja liiketoiminnan ja Data ja tekoäly-yksikön välille.

Tutkimus toteutettiin toimintatutkimuksena, jonka alussa tutustuttiin teoriaan ja alan parhaisiin käytäntöihin datan laadun ja sen hallinnan osalta. Näihin aikoihin myös tunnistettiin liiketoimintavaatimukset ja halutut toiminnallisuudet laadunhallintamenetelmien osalta, joiden jälkeen toiminnallisuuksia työstettiin ja otettiin käyttöön data-alustalla. Tämän jälkeen toteutuksen toimivuutta seurattiin ja toimintamalleja kehitettiin tarvittaessa. Tuloksia reflektoitiin ja arvioitiin jatkuvasti, jotta lopputulos vastaisi haluttua tavoitetilaa.

Tutkimuksen lopputuloksena halutut datan laadunhallinnan menetelmät ja periaatteet saatiin käyttöönotettua data-alustalle sekä salkkuennusteprosessiin, muutamia poikkeuksia lukuun ottamatta. Jatkossa moduulia ja sen toiminnallisuuksia voidaan laajentaa muihinkin prosesseihin ja niiden käyttämään dataan. Yhtenä potentiaalisista kehitysalueista datan laadunhallinnassa on tekoäly ja koneoppiminen. Näiden hyödyntäminen datan laadunhallinnassa ja siihen liittyvien akateemisten tutkimusten määrän uskotaan lisääntyvän tulevaisuudessa. Myös vastaavien ratkaisujen käyttöönottoa tullaan mahdollisesti tekemään Helenillä jossain vaiheessa.

ABSTRACT

Lappeenranta–Lahti University of Technology LUT

School of Engineering Science

Industrial Engineering and Management

Aino Solin

Implementation of data quality management at an energy company

Master's thesis

2022

76 pages, 21 figures, 12 tables and 1 appendix

Examiners: Professor Timo Kärri, Postdoctoral Researcher Lasse Metso

Keywords: Data quality management, data quality, implementation

This thesis investigated how data quality management could be implemented at Helen Oy both in their data platform and in electricity price hedging process. The study was preceded by a current state analysis of data quality and its management carried out in 2021, where deficiencies were found, for example, in the quality of data used in the portfolio forecasting process, such as contract data. The goals and limitations of the work were drawn up based on the deficiencies identified in the current state analysis and reference framework of the selected supplier. The goal was to implement data quality management methods on the data platform and to develop functional data quality management cooperation and operating models between business operations and the Data and Artificial Intelligence unit.

The research was carried out as action research, which began with getting to know the theory and the best practices in data quality and its management. Around this time, the business requirements and desired functionalities were also identified in terms of quality management methods, after which the functionalities were worked on and implemented on the data platform. After this, the functionality of the implementation was monitored, and operating models were improved. The results were constantly reflected and evaluated so that the results corresponded to the desired target state.

As a result of the research, the desired data quality management methods and principles were implemented in the data platform and in the electricity price hedging process, with a few exceptions. In the future, the data quality module and its functionalities can be extended to other processes and their data. One of the potential areas of development in data quality management lie in artificial intelligence and machine learning. It is believed that the utilization of these in data quality management and the number of related academic studies will increase in the future. The implementation of corresponding solutions will also possibly be done at Helen at some stage.

KIITOKSET

Tahdon kiittää Heleniä mahdollisuudesta tehdä töitä mahtavien osaajien ja asiantuntijoiden keskellä mielenkiintoisten ja innostavien aiheiden parissa. Erityisesti haluan kiittää esihenkilöäni, keneltä olen saanut arvokkaita oppeja ja ohjausta koko tutkimusprojektin ajan. Lisäksi haluan kiittää kaikkia tutkimuksessa mukana olleita kollegoita, jotka mahdollistivat tutkimuksen onnistumisen.

Iso kiitos myös ystävilleni ja läheisilleni, jotka tukivat ja kannustivat minua opintojeni kaikissa vaiheissa.

21.11.2022

Aino Solin

Lyhenneluettelo

DAMA	Data Management Association
DIRA	Digitaaliset Ratkaisut
DMAIC	Define, Measure, Analyze, Improve and Control - määritä, mittaa, analysoi, kehitä, ohjaa
DQ	Data Quality - datan laatu
DQAF	Data Quality Assessment Framework - datan laadun arviointiviitekehys
DQM	Data Quality Management - datan laadunhallinta
EHT	Energianhallinta ja tukkukauppa
ISO	International Organization for Standardization
MAP	Myynti ja asiakaspalvelu
RACI	Responsible, Accountable, Consulted, Informed - vastuumääräysmatriisi
RCA	Root Cause Analysis - juurisyyanalyysi
TDQM	Total Data Quality Management - kokonaisvaltainen datan laadunhallinta
TIQM	Total Information Quality Management - kokonaisvaltainen informaation laadunhallinta

Sisällysluettelo

1 Johdanto

Data- ja analytiikka ovat keskeisessä roolissa asiakassuhteen ja asiakaskokemuksen kehityksessä, sillä niiden avulla voidaan kehittää palveluja vastaamaan asiakkaiden tarpeisiin yhä paremmin sekä parantaa liiketoimintojen tuloksellisuutta. Data ja teknologia kuitenkin muuttuu ja kehittyy alati, joten myös sen hallitsemiseksi on kehitettävä menetelmiä, toimintatapoja ja käytettävä oikeanlaisia työkaluja. Erityisesti liiketoimintakriittisten prosessien käyttämästä datasta on pidettävä huolta, jotta voidaan välttyä tulonmenetyksiltä.

Tässä tutkimuksessa käsitellään datan laadunhallinnan käyttöönottoa Helenin data-alustalla sekä salkkuennusteprosessissa, jonka avulla suojataan sähkön hintaa. Salkunhoito on kriittinen riskienhallinnan toiminto, jonka tarkoituksena on hallita markkinariskejä, vakauttaa kassavirta ja varmistaa riittävä kate. Viime vuosina sähkön hinta on noussut jyrkästi johtuen geopoliittisista muutoksista maailmalla, sääolosuhteista ja muista tekijöistä. Osittain näistä syistä salkkuennusteeseen ja sen laatuun on nyt Helenissä kiinnitetty erityisen paljon huomiota, vaikka se on maailmantilanteestakin riippumatta kriittinen liiketoimintaprosessi. Suojautumalla hintavaihtelulta voidaan luoda edellytykset tasaiselle ja vakaalle tuloksentekokyvylle. Mikäli sähkö alihinnoitellaan, eli suojataan väärin, sillä on suora vaikutus kannattavuuteen, sillä sähköä on myyty liian pienellä katteella ja tulos heikkenee. Virheellisellä, epätarkalla tai myöhässä tulleella datalla voi siis olla taloudellisesti merkittäviä suoria vaikutuksia. Etenkin lyhyellä aikavälillä ennusteiden tarkkuuden paraneminen voi vaikuttaa merkittävästi riskiin ja tulokseen, kun taas pitkällä aikavälillä riski tasoittuu.

Tutkimusta edelsi vuoden 2021 kesällä toteutettu nykytila-analyysin datan laadusta, eli siitä missä datan laadunhallinnan ja laadun edistämisen saralla ollaan. Analyysin tuloksista selvisi, että uusia ja etenkin vanhoja tietolähteitä integroitaessa uudelle data-alustalle, Helenin Data ja tekoäly-yksikkö oli havainnut erilaisia puutteita datan laadussa, kuten sopimusmäärien eroja raportointikannan ja lähdejärjestelmän välillä sekä datan saatavuudessa. Analytiikan hyödyllisyys päätöksenteossa on pitkälti riippuvainen datan laadusta, joten laadun jatkuva seuranta, kehittäminen ja virheiden ennaltaehkäiseminen on ehdottoman tärkeää. Data-alustalla ei siis aiemmin ollut keskitettyä tapaa hallita tai monitoroida datan laatua, vaan sitä

oli toteutettu pistemäisesti ja erilaisilla työkaluilla yksittäisten henkilöiden toimesta omiin tarpeisiin. Näistä syistä datan laadunhallintamenetelmille on todettu olevan tarve.

1.1 Työn tavoitteet

Työn tavoitteena on datan laadunhallintamenetelmien käyttöönoton liittyvän toimintatutkimuksen avulla löytää vastaus päätutkimuskysymykseen:

- Miten datan laadunhallinnan menetelmät ja periaatteet voidaan ottaa käyttöön Data ja tekoäly -yksikön prosesseissa ja pienasiakasmyynnin salkkuennusteprosessissa?

Päätutkimuskysymyksen tueksi muodostettiin myös seuraavat aututkimuskysymykset:

- Miten datan laadunhallintaa hyödynnetään?
- Mitä datan laadun mittareita data-alustan ja salkkuennusteen osalta tulisi olla?

Tutkimuksessa selvitettiin kuinka datan laadunhallinnan käytäntöjä ja periaatteita voidaan ottaa käyttöön Data ja tekoäly-yksikön prosesseihin sekä salkkuennusteprosessiin. Tässä hyödynnettiin teoriaa datan laadunhallinnan viitekehyksistä, jotta saatiin kuvaa mitä perusperiaatteita, aktiviteetteja ja tehtäviä kyseisten viitekehysten käyttöönotossa on. Tavoitteena oli jatkojalostaa laadunhallinnan viitekehystä Helenin toimintaympäristöön sopivaksi sekä jalkauttaa periaatteet ja työkalut data-alustalle sekä liiketoimintaprosessiin. Tämä kattoi siis salkkuennusteprosessissa käytetyn datan sisällön sekä dataputkien teknisten mittareiden määrittämisen, mittaamisen ja analysoimisen. Datan laadun mittariston määrityksen tueksi löytyi paljon teoriaa datan laadun dimensioista sekä siitä, miten niitä mitataan. Kysymykseen ”miten datan laadunhallintaa hyödynnetään?”, eli minkälaisia menetelmiä datan laadunhallintaan on sekä mitä arvoa tuo toiminto tuottaa, löydettiin vastauksia datan laadunhallinnan viitekehysten ja heikon datan laadun vaikutusten- ja kustannusten teoriasta.

1.2 Työn rajaus

Työn tarkoituksena ei ollut suunnitella ja ottaa käyttöön kaiken kattavaa datan laadun viitekehystä, joka sopeutuisi kaikkiin käyttötarkoituksiin, vaan ottaa käyttöön sellaisia perustavanlaatuisia laadunhallintamenetelmiä, joita voidaan hyödyntää nykytila-analyysissa ja

vaatimusten kartoituksessa esille nousseiden asioiden korjaamiseen tai kehittämiseen. Lisäksi yhtenä rajauksena oli, että rajataan tarkastelu jo käytössä tai kokeilussa oleviin tai työkaluihin, eli ettei työn piirissä tutkittaisi uusia työvälineitä. Näihin rajauksiin päädyttiin, jottei työn laajuus kasva liian suureksi ja jotta halutut muutokset saadaan varmasti käyttöönotettua. Datan laatu ja sen hallinta ovat laajempia käsitteitä kuin tutkimuksen aikataulu antaa periksi, jonka vuoksi kaikkia näkökulmia tai osa-alueita aiheesta ei voitu ottaa huomioon.

1.3 Tutkimuksen toteutus

Tutkimusmenetelmäksi valittiin toimintatutkimus, sillä tutkimuksen tekijä toimi aktiivisesti projektissa mukana havainnoinnin sijaan sekä hyödynsi aiempaa työkokemustaan ja ymmärrystä toimeksiantajansa datan laadun nykytilasta (Juuti & Puusa 2020). Pääsääntöisenä empiirisenä aineistona toimivat työpajojen sekä haastattelujen tulokset sekä valitun toteutuskumppanin laatima laadunhallinnan viitekehys, jota muokattiin kirjallisuuden, haastattelujen ja työpajojen avulla sopivaksi Helenin toimintaympäristöön. Toimintatutkimus toteutettiin syklisenä mallina, joka sisälsi neljä vaihetta. Jokaisen syklin päätteeksi reflektoinnin tuloksia hyödynnettiin seuraavan vaiheen tukena.

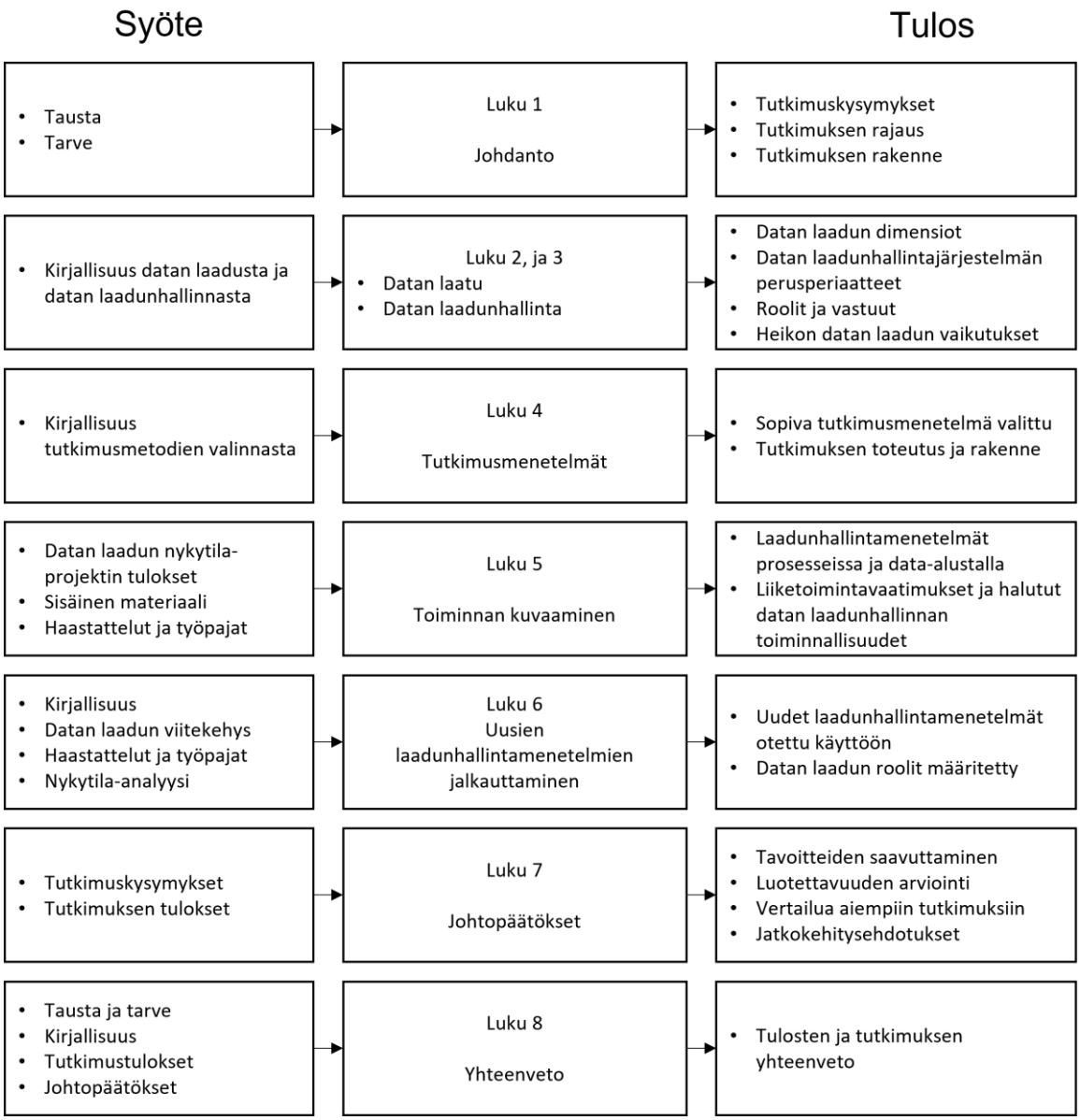
Ensimmäiseen sykliin kuului nykytilan ja vaatimusten kartoitusta, toisessa syklissä keskityttiin pääasiassa teknisen arkkitehtuurin työstöön ja työkalujen asennukseen sekä mittarien määritykseen, kolmannessa datan laatua analysoitiin ja laadittiin juurisyyanalyyseja virheiden synnystä, ja viimeisessä syklissä pyrittiin korjaamaan esimerkkicase-ongelma sekä laatimaan toimintaohjeet sekä tukidokumentaatiota. Ensimmäisen ja toisen vaiheen jälkeen vaatimuksia uudelleenarvioitiin ja määrittelyjä tehtiin osittain uudelleen, joka johti joistain toiminnallisuuksista luopumiseen. Lopulta työ johti laadunhallintaviitekehyksen jalostamiseen sekä sen käyttöönottoon salkkuennusteprosessissa sekä data-alustalla.

Haastateltavat ja työpajojen osallistujat valittiin lumipallomenetelmällä (Tuomi & Sarajärvi 2019), eli yksikön johtaja sekä prosessista vastaava asiantuntija nimesi listan henkilöistä, joilla uskoi olevan riittävästi tietoa tutkittavasta aiheesta. Listaan pyrittiin tunnistamaan sellaiset henkilöt, jotka ovat mukana läpi prosessin ja / tai vastaavat laadunvarmennuksesta jossakin tietyssä prosessin vaiheessa, eli heillä on riittävää kokemusta ja tietämystä käyttämistään menetelmistä omassa prosessin vaiheessaan. Empiiristen aineistojen kerääminen toteutettiin eri vaiheissa välillä maaliskuu-lokakuu 2022. Osallisten osaamistaso, roolit ja

vastuualueet poikkesivat toisistaan jonkin verran, joten työpajat ja haastattelut toteutettiin teemahaastatteluina, ilman varsinaista haastattelukysymyspohjaa sekä suurimmilta osin Teamsin välityksellä.

1.4 Tutkimuksen rakenne

Tutkimus on jaettu kahdeksaan päälukuun. Raportin rakenne ja jokaisen vaiheen keskeiset syötteet ja tulokset ovat esitetty kuvassa 1 alla.



Kuva 1: Tutkimuksen rakenne

Luvussa 1 tutustutetaan lukija aiheeseen ja esitetään tutkimuskysymykset, tavoitteet, rajaus ja toteutustapa. Luvuissa 2 ja 3 käydään läpi olemassa olevaa kirjallisuutta ja tutkimuksia liittyen datan laatuun ja datan laadunhallintaan. Näiden avulla löydetään vastauksia esitettyihin tutkimuskysymyksiin sekä tunnistetaan datan laadunhallinnan pääperiaatteet, joiden mukaan tutkimuksessa edetään. Neljännessä luvussa esitetään tutkimuksessa käytetty tutkimusmenetelmä. Kappale 5 keskittyy toiminnan kuvaamiseen, eli siinä käydään läpi nykytilannetta ja vaatimuksia sekä toiveita koskien uusia laadunhallintamenetelmiä. Kuudennessa kappaleessa käydään läpi, kuinka syklien aikana edettiin ja mitä uusia toiminnallisuuksia ja toimintatapoja saatiin otettua käyttöön. Luvussa 7 esitetään johtopäätökset tutkimuksesta, tuloksista ja arvioidaan niiden luotettavuutta, sekä vertaillaan työtä aiempaan tutkimukseen. Viimeisessä luvussa muodostetaan yhteenveto tutkimuksesta.

2 Datan laatu

Datasta usein puhutaan tietojärjestelmien yhteydessä ja informaatiosta sen jatkojalostuksen tuotteena. DAMA (Data Management Association Data Management) määrittelee *datan* jonain uudelleentulkittavissa olevana muodollisena tietona, joka soveltuu kommunikoitavaksi, tulkittavaksi tai käsiteltäväksi (DAMA n.d.). Data on siis faktojen kuvaamista erilaisissa muodoissa, kuten numeroina, tekstinä, kuvina, äänenä tai videona (Mosley et al. 2010, s. 2), jota voi syntyä muun muassa liiketoimintaprosesseista, käyttäjän syötteestä, sensoreista tai esimerkiksi mittauslaitteista. Toisin sanoen, data on jonkinlainen tiedonesitys, jolla on potentiaalista arvoa. Sitä jatkojalostamalla saadaan *informaatiota*, eli tietoa, joka on muokattu käyttäjän näkökulmasta hyödylliseen muotoon. (Baskarada & Koronios 2013, s. 7)

Informaation voidaan tulkita olevan myös jollain tavalla strukturoitua ja järjesteltyä, kun taas data on luonteeltaan järjestelemätöntä ja käsittelemätöntä, eli dataa sellaisenaan voidaan pitää malleina, joilla ei ole vielä merkitystä ja informaatiota tulkittuna datana, jolla on merkitys (Baskarada & Koronios 2013, s. 7). Informaatio on dataa kontekstissa ja data ilman kontekstia on merkityksetöntä (Mosley et al. 2010, s. 2).

Tiedolla tai tietämyksellä (knowledge) usein tarkoitetaan datan ja informaation ylempää käsitettä, mikä viittaa jatkojalostettuun informaatioon, jonka avulla voidaan välittää ymmärrystä, kokemusta ja asiantuntemusta, kun niitä sovelletaan tiettyyn ongelmaan tai toimintaan. Viisaus (wisdom) nähdään kerrytetyksi tiedoksi, jota hyödyntämällä voidaan soveltaa tuota opittua tietoa uusiin ongelmiin tai tilanteisiin. (Finto 2018; Baskarada & Koronios 2013, s. 7)

Myös laadulle esitetään kirjallisuudessa erilaisia määritelmiä, mutta esimeriksi ISO:n mukaan laatu määritellään sellaiseksi missä määrin jonkin asian luontaiset ominaisuudet täyttävät sille asetetut vaatimuksensa (International Organisation of Standardization 2015, kappale 3.6). Toisin sanoen esimerkiksi organisaation sisäisesti voi olla määritetty yhteisesti jokin standardi, laadulliset luokat tai raja-arvot, jonka tuotteen tai palvelun tulee täyttää, että sen laadun voidaan määritellä olevan ”hyvä”, ”huono” tai jotain siltä väliltä. Toisaalta laatu on myös subjektiivista ja asiakkaan määrittelemää, eli asiakkaalla voi olla ennakko-odotuksia saamastaan palvelun tai tuotteen laadusta, joka voi riippua muun muassa aiemmista kokemuksista tai kilpailijoiden tarjoaman laadusta. Tästä syystä laatua määritellään usein myös

käyttökelpoisuuden (fitness-for-use) tai tarkoitukseen sopivuuden (fitness-for-purpose) kautta, jossa asiakas itse määrittelee sopivuuden (fitness). (Kiran 2019)

Näiden määritelmien kautta *datan laatu* voidaan määritellä sellaiseksi, että missä määrin asiakas, eli datan käyttäjä kokee datan olevan käyttökelpoista tai tarkoitukseensa sopivaa (Flores & Sun 2018, s. 2), tai missä määrin datan ominaisuudet vastaavat oletettuja tarpeita tietyissä olosuhteissa (International Organisation of Standardization 2008, kappale 4.3). Esimerkiksi mikäli myyjän havainto on, että myyntiraportilta puuttuu dataa tai numerot eivät vastaa todellisuutta, voi hän kokea datan huonolaatuiseksi, sillä hän tuntee datan sisällön hyvin, eli mitä asioita datasta hän tarvitsee ja mihin hän niitä tarvitsee. Datan jalostusprosessin eri vaiheissa on mukana useita rooleja, jotka kukin voivat tunnistaa oman työn kannalta oleellisia puutteita datasta. Rooleja avataan lisää kappaleessa 3.3.

Virheellistä dataa on siis se, joka ei täytä sille määritettyjä vaatimuksia. Datassa voi esiintyä poikkeamia (outlier / anomaly), eli datapisteitä, jotka eroavat huomattavasti tyypillisestä datan luonteesta (D'Urso, 2016, s. 2). Numeerisessa datassa nämä voivat näkyä erillisinä pisteinä erossa muista datapisteistä, siinä missä muut noudattavat jakaumaa tai ne kuuluvat luontaisiin klustereihinsa (MIT Critical Data, 2016, s. 165). Poikkeamat eivät kuitenkaan aina ole merkkejä virheellisestä datasta, mutta niiden syitä on tarpeen tutkia, sillä ne voivat selittää ilmiön tai datan piilevää luonnetta (MIT Critical Data, 2016, s. 163). Heikon datan laadun juurisyitä voivat olla muun muassa sama tieto useassa lähteessä, epäkunnossa olevat laitteet, subjektiivinen arviointi tietojen syöttämisessä (ei koeta jonkin tiedon olevan tärkeää ja jätetään se kirjaamatta), rajallinen prosessointikapasiteetti, tasapaino turvallisuuden ja saavutettavuuden välillä, epäyhtenäinen terminologia tietojärjestelmien ja organisaatioyksiköiden välillä, datan määrä ja erityyppiset formaatit, tiedon kirjaussääntöjen tiukkuus tai niukkuus, tai muuttuvat datatarpeet. (MIT Critical Data, 2016, s. 163; Lee et al. 2006, s. 81–82)

2.1 Datan laadun dimensiot

Informaatio ja data -käsitteiden eroista huolimatta, näitä termejä käytetään usein rinnakkain (Flores & Sun 2018, s. 2), sillä laadun näkökulmasta niillä on hyvin samankaltaisia ominaisuuksia kuten ajankohtaisuus (onko data tai informaatio ajankohtaista vai vanhentunutta), täydellisyys (puuttuuko dataa tai informaatiota) ja tarkkuus (vastaako data tai informaatio

todellisuutta). Selkeyden vuoksi seuraavissa kappaleissa puhutaan datan laadusta, mutta samankaltaisia jaotteluita voidaan käyttää myös informaation ja osittain myös tietämyksen osalta. Huomattakoon vielä, että usein datan laadusta puhuttaessa viitataan niin sanottuihin teknisiin ongelmiin ja taas informaation laatu käsittää niin sanotut ei-tekniset tai sisällölliset ongelmat (Zhu et al. 2014).

Jotta datan laatua voidaan määritellä, sille täytyy asettaa tiettyjä vaatimuksia. Datan laadun määritelmän mukaan ”hyvälaatuisen” datan tulee olla käyttökelpoista tai tarpeeseen sopivaa. Useimmat liiketoimintaprosessien käyttämät sovellukset ovat riippuvaisia datasta, jonka tulee täyttää tietyt vaatimukset, jotta prosessi tuottaa haluttuja lopputuloksia. Nämä vaatimukset voivat olla esimerkiksi liiketoimintayksikön itse määrittelemiä liiketoimintasääntöjä, järjestelmään sisäänrakennettuja, tai alalla vallitsevia parhaita käytäntöjä ja standardeja, sekä dataformaattien eheyden varmistamista eri järjestelmien tiedonsiirron välillä. (Mosley et al. 2010, s. 295–296)

Laadullisia vaatimuksia pohtiessa voidaan dataa tarkastella eri näkökulmista, joista usein puhutaan datan laadun dimensioina tai ominaisuuksina, joista osa on datalle luontaisia ja osa järjestelmästä riippuvaisia. Näitä ominaisuuksia on kuvattu taulukkoon 1 alla.

Taulukko 1: Datan laadun ominaisuudet (ISO/IEC 2020, s. 11)

Ominaisuudet	**Datan laatu**	
	Luontainen	Järjestelmästä riippuvainen
Tarkkuus (Accuracy)	X	
Täydellisyys (Completeness)	X	
Johdonmukaisuus (Consistency)	X	
Uskottavuus (Credibility)	X	
Ajankohtaisuus (Currentness)	X	
Esteettömyys (Accessibility)	X	X
Vaatimustenmukaisuus (Compliance)	X	X
Luottamuksellisuus (Confidentiality)	X	X
Tehokkuus (Efficiency)	X	X
Täsmällisyys (Precision)	X	X
Jäljitettävyys (Traceability)	X	X
Ymmärrettävyys (Understandability)	X	X
Saatavuus (Availability)		X
Siirrettävyys (Portability)		X
Palautettavuus (Recoverability)		X

Datalle luontaisia laadullisia ominaisuuksia ovat tarkkuus, täydellisyys, johdonmukaisuus, uskottavuus, ajankohtaisuus, esteettömyys, vaatimustenmukaisuus, luottamuksellisuus, tehokkuus, täsmällisyys, jäljitettävyys sekä ymmärrettävyys. Järjestelmäriippuvaisia ominaisuuksia laadun näkökulmasta ovat esteettömyys, vaatimustenmukaisuus, luottamuksellisuus, tehokkuus, täsmällisyys, jäljitettävyys sekä ymmärrettävyys, saatavuus, siirrettävyys sekä palautettavuus. Osa ominaisuuksista ovat sekä luontaisia että järjestelmästä riippuvaisia.

Luontaisilla ominaisuuksilla viitataan ISO/IEC:n mukaan siihen, missä määrin datalla on luontainen potentiaali täyttää ilmoitettuja ja oletettuja tarpeita, kun dataa käytetään tietyissä olosuhteissa. Luontainen potentiaali viittaa suoraan dataan, eli esimerkiksi tämän kyseisen data-alueen saamiin arvoihin sekä mahdollisiin rajoitteisiin (kuten liiketoimintasääntöihin, jotka säätelevät laatua tietyssä sovelluksessa), tai tietoarvojen suhteisiin ja metatietoihin. Järjestelmästä riippuvaiset ominaisuudet tarkoittavat sellaisia, missä määrin datan laatu

saavutetaan ja säilytetään tietojärjestelmässä, kun dataa käytetään tietyissä olosuhteissa. Eli toisin sanoen datan laatu voidaan nähdä olevan riippuvainen käytettävästä teknologiasta. Laatu varmistetaan erilaisilla komponenteilla, joita voivat olla muun muassa fyysiset laitteet (joilla varmistetaan datan tallennus ja saavutettavuus, kuten fyysinen tietovarasto, tai datan tarkkuus, esimerkiksi mittauslaitteet) tai ohjelmistot (esimerkiksi varmuuskopio-ohjelmistot, joilla varmistetaan palautettavuus, taikka apuohjelmat, millä voidaan visualisoida datan kulku eri järjestelmien välillä varmistaakseen jäljitettävyyden), sekä muut ohjelmistot (kuten integraatio / migraatiotyökalut, joilla mahdollistetaan datan siirrettävyys). (ISO/IEC 2020, s. 10)

Massadataa (big data) muodostuu valtavia määriä ja ennennäkemättömällä nopeudella, minkä lisäksi se on käsiteltävä tietyn ajan sisällä. Datatyyppien vaihtelu on myös suurta, eli erilaisia tietotyyppejä, mitkä tarvitsevat erityistä ja kompleksisempaa käsittelyä, on paljon. Arkkitehtuuri tällaisen tiedon käsittelyyn on usein myös monimutkainen, joka vaikeuttaa datan jäljitettävyyttä. Nämä seikat tekevät erityisesti massadatan laadun mittaamisesta haastavampaa, jonka vuoksi massadatan laadulle esitetään kirjallisuudessa omia jaotteluja. Taulukkoon 2 alla on kerätty näitä jaotteluja ja missä lähteessä niitä käsitellään.

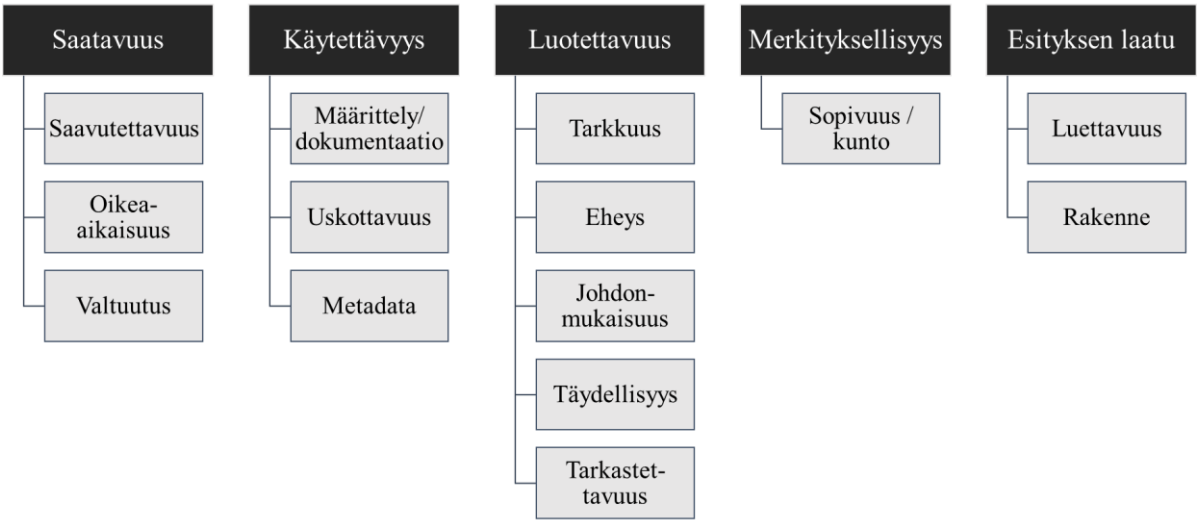
Taulukko 2: Massadatan laadun jaotteluja

Lähde	Jaottelu	Kategoriat / ulottuvuudet
Cai & Zhu, 2015	5 kategoriaa, 14 ulottuvuutta	Saatavuus: saavutettavuus, oikea-aikaisuus, valtuutus Käytettävyys: määrittely/dokumentaatio, uskottavuus, metadata Luotettavuus: tarkkuus, eheys, johdonmukaisuus, täydellisyys, tarkastettavuus Merkityksellisyys: sopivuus/kunto Esityksen laatu: luettavuus, rakenne
Batini et al., 2015	7 kategoriaa, 17 ulottuvuutta	Tarkkuus: oikeellisuus, kelpoisuus, tarkkuus Täydellisyys: asianmukaisuus, relevanssi Redundanssi: minimaalisuus, kompaktius, ytimekkyys Luettavuus: ymmärrettävyys, selkeys, yksinkertaisuus Saavutettavuus: saatavuus Johdonmukaisuus: yhtenäisyys, johdonmukaisuus Luottamus: uskottavuus, luotettavuus, hyvämaineisuus
Taleb et al., 2018	4 kategoriaa, 18 ulottuvuutta	Sisäinen: tarkkuus, oikea-aikaisuus, johdonmukaisuus, täydellisyys Kontekstuaalinen: maine, osuvuus, saatavuus, määrä, lisäarvo, uskottavuus Edustavuus: tulkittavuus, edustava, ytimekkyys edustus, johdonmukaisuus käsiteltävyys, ymmärrettävyys Esteettömyys: pääsy, turvallisuus
Wook et al., 2021	4 kategoriaa, 16 ulottuvuutta	Sisäinen: Tarkkuus, objektiivisuus, uskottavuus, maine Kontekstuaalinen: arvoa tuottava, merkityksellisyys, ajankohtaisuus, täydellisyys, oikea volyymi Edustavuus: tulkittavuus, ymmärrettävyys, ytimekäs esitys, johdonmukainen esitys Esteettömyys: saatavuus, pääsyn turvallisuus, käytön helppous

Wook et al. (2021) sekä Taleb et al. (2018) lähteissä esitetyissä kategorisissa jaotteluissa toistuvat neljä pääkategoriaa: sisäinen, kontekstuaalinen, edustavuus sekä esteettömyys. Ne eivät ole kuitenkaan ainoastaan massadatalle ominaisia kategorioita tai ulottuvuuksia, vaan niitä on myös käsitelty yleisinä datan laadun ominaisuuksina (Sebastian-Coleman, 2013, liite B; Wang & Strong, 1996, s. 20). Sisäiset ominaisuudet viittaavat datan luontaisiin

ominaisuuksiin tai sen sisältöön, jonka alle luetaan muun muassa tarkkuus, johdonmukaisuus, oikeellisuus, kelpoisuus (Taleb et al., 2018, s. 4). Kontekstuaaliset datan laadun ominaisuudet ovat liitoksissa ja riippuvat kyseessä olevasta tehtävästä, eli kuinka hyvin data vastaa käyttäjän odotuksia tai kykenee tuottamaan käyttäjälleen arvoa (Wook et al. 2021, s. 3). Edustavuudella tarkoitetaan datan laatua suhteessa esitystapaan, toisin sanoen kuinka ymmärrettävää ja helppokäyttöistä data on, tai kuinka johdonmukaisesti tai ytimekkäästi se on esitetty (Sebastian-Coleman, 2013, liite B). Esteettömyys puolestaan korostaa järjestelmäkohtaisia saavutettavuutta parantavia ominaisuuksia, joihin luetaan esimerkiksi pääsyn turvallisuus (Wook et al. 2021, s. 3).

Cai & Zhu (2015) laativat hyvin samankaltaisen jaottelun (kuvassa 2), mutta missä massadatan laadun ominaisuudet on jaettu viiteen yläluokkaan, joiden alle on tunnistettu 1–5 ulottuvuutta. Neljää ensimmäistä luokkaa nähdään laadulle luontaisina ominaisuuksina, kun taas viimeinen luokka (esityksen laatu ja sen alla olevat ulottuvuudet) on ikään kuin lisäominaisuus, mikä parantaa asiakastyytyväisyyttä.



Kuva 2: Massadatanlaadun ominaisuudet (mukaillen Cai & Zhu, 2015)

Datan saatavuudella viitataan tietojen hankkimisen käyttömukavuusasteeseen, mikä on jaettu kolmeen ulottuvuuteen: saavutettavuus, oikea-aikaisuus (tai ajantasaisuus) ja valtuutus (onko käyttäjällä oikeutta dataan). Käytettävyyden käsite tarkoittaa sitä, ovatko tiedot hyödyllisiä ja täyttävätkö ne käyttäjien tarpeet. Käytettävyys voidaan jakaa kolmeen alaulottuvuuteen, jotka ovat määrittely/dokumentaatio, uskottavuus ja metadata. Luotettavuus koostuu viidestä ulottuvuudesta, jotka ovat: tarkkuus, eheys, johdonmukaisuus, täydellisyys, sekä

tarkastettavuus. Merkityksellisyys tarkoittaa datan ja käyttäjän odotusten tai vaatimusten välistä korrelaatiota, eli toisin sanoen kuinka hyvin data sopii käyttäjänsä tarkoitukseen. Esityksen laatu taas viittaa siihen, kuinka dataa kuvataan ja miten kuvausmenetelmä auttaa käyttäjiä ymmärtämään dataa paremmin. Eli mitä selkeämmin data on kuvattu, luokiteltu ja on ylipäätään selkeästi ymmärrettävissä, niin sitä parempi luettavuus datalla on. Rakenne viittaa suoraan tietorakenteeseen, eli esimerkiksi siihen, kuinka hankalaa puolistrukturoitua tai strukturoimatonta dataa on muuttaa strukturoituun muotoon. (Cai & Zhu, 2015)

Batini et al. (2015) ovat jaotelleet ominaisuuksien yläkategoriat hieman eri tavalla, mutta sisältäen samoja ulottuvuuksia kuin muissakin lähteissä. Ne noudattavat kuitenkin samanlaista kaavaa kuin Wook et al. (2021) sekä Taleb et al. (2018), eli datan sisältöön viittaavia kategorioita ovat esimerkiksi tarkkuus ja täydellisyys, kun taas kontekstuaaliseen ulottuvuuteen voitaisiin niputtaa yhteen esimerkiksi luottamus ja redundanssi. Edustavuuteen voitaisiin lukea mukaan luettavuus sekä johdonmukaisuus ja esteettömyyteen saavutettavuus sekä luottamus, sillä Batini et al. käsittelevät turvallisuutta jälkimmäisen ulottuvuuden yhteydessä.

2.1.1 Tarkkuus

Datan tarkkuus (accuracy) tyypillisesti tarkoittaa sitä, kuinka lähellä ”todellisia” arvoja data on (ISO/IEC 2020, s. 11; Batini & Scannapieco, 2016, s. 23). Tarkkuutta voidaan myös tarkastella eri näkökulmista, kuten strukturaalinen tarkkuus ja temporaalinen tarkkuus. Strukturaalista tarkkuutta tarkoitetaan, kun data-arvoa voi tarkastella tietyn aikajakson sisällä ja arvon ei odoteta muuttuvan, kun taas temporaalista tarkkuutta mitataan esimerkiksi ajankohtaisuuden (currentness), volatiliteetin (volatility) ja ajantasaisuuden (timeliness) avulla (Batini & Scannapieco, 2016, s. 27).

Strukturaalisesta datan tarkkuudesta esimerkkinä voi toimia muun muassa nimikenttien arvot, eli ovatko nimet esitetty oikein. Jos asiakkaan nimi on todellisuudessa ”Laura”, mutta tietokannassa lukee ”Lura”, voidaan todeta, ettei strukturaalinen tarkkuus ole riittävällä tasolla. Mikäli taas datan tulisi olla ajankohtaista, mutta tieto ei ole päivittynyt vastaamaan todellisuutta, temporaalinen tarkkuus voidaan nähdä huonoksi. (Batini & Scannapieco, 2016, s. 23–24). Joissain tapauksissa datan tarkkuus voi olla myös hyvin yksinkertaisesti mitattavissa, esimerkiksi henkilön syntymäpäivän tulisi olla oikeasti olemassa oleva

päivämäärä, eli ei 31.2., koska helmikuussa on tunnetusti vain 28 päivää (paitsi karkausvuosina). Tai mikäli syntymäpäivä on ilmoitettu väärässä formaatissa, esimerkiksi Yhdysvaltain päivämäärämuodossa eurooppalaisen tyylin sijaan, niin datan tarkastelija ei voi tietää onko se väärin lähtökohtaisesti vai ainoastaan epähuomiossa kirjattu väärin. Kummassakaan tapauksessa arvo ei ole riittävän tarkka, mikäli arvon käyttäjä ei osaa varmuudella sanoa mikä se on. (Olson 2008, s. 29)

Tarkkuutta voidaan mitata esimerkiksi laskemalla, kuinka monta virhettä datassa on. Tarkkuuden laskemiseksi voidaan olettaa esimerkiksi, että 100 % on haluttu tarkkuus, josta vähennetään virheellisten rivien osuus kokonaisrivimäärästä. Virheellisten rivien määrittämiseksi täytyy kuitenkin olla jonkinlaisia sääntöjä kuvaamassa, mikä data lasketaan virheelliseksi. Sääntö voi olla vaikkapa, että henkilötunnuksen tai y-tunnuksen täytyy noudattaa tiettyä kaavaa, joka on jo tiedossa. Kaava voidaan laskea muun muassa yksinkertaisella SQL-haulla, eli valitsemalla ensin kaikki rivit, mitkä eivät vastaa määritettyjä oletuksia ja jakamalla ne kaikkien taulussa olevien rivien määrällä, mikä vähennetään halutusta tarkkuudesta (100 %), jolloin saadaan kentän prosentuaalinen tarkkuus. (Lee et al. 2006, s.55–60) Syntymäpäivän päivämäärän oikeellisuutta voidaan tarkistaa esimerkiksi SQL-kielen ISDATE()-funktiolla, mikä tarkistaa onko päivämäärä validi ja palauttaa arvon 0, mikäli se ei ole ja vastaavasti palauttaa arvon 1, jos on (Refsnes Data, n.d.).

Tarkkuutta ei pystytä aina todistamaan oikeaksi, sillä sen oletuksena on, että arvot vastaavat tosielämän arvoja, joita voi olla haastava varmentaa ”oikeiksi”, mikäli poikkeamat ovat hyvin pieniä tai mikäli ”oikeaa” arvoa ei ole saatavilla (McGilvray 2008, s. 31–32). Tästä syystä usein oikeellisuutta / validiteettia (validity) on usein helpompaa mitata, sillä sen oletuksena on, että arvot ovat yhdenmukaisia määritetyllä alueella. Toisaalta, mikäli kenttä on validi, eli vaikkapa postinumero on teknisesti oikein ja olemassa oleva, mutta sitä on käytetty väärän postinumeroalueen kanssa, ei data tällöin läpäise tarkkuusvaatimuksia. (Kaushik 2020)

2.1.2 Oikeellisuus / Validiteetti

Datan oikeellisuudella (validity) tarkoitetaan, että data tallennetaan, muutetaan tai esitetään tietoalueelle tyypillisessä muodossa yhdenmukaisesti. Odotuksena on myös, että data on yhdenmukaista muiden vastaavien attribuuttiarvojen kanssa. Tämä edellyttää, että jokaiselle

kentälle on määritetty ”oikeellisuus”, eli esimerkiksi jokin tietty kaava tai formaatti, jota datan pitää noudattaa ja oikeellisuus vaihtelee tyypillisesti kentittäin. (Mosley et al. 2010, s. 297)

Oikeellisuustarkistuksia voi olla kaavan ja formaatin lisäksi myös liiketoimintasäännöt, datan syöttöstandardit, sekä minimi ja maksimirajat datalle. Mikäli kentän tulisi sisältää vain numeroarvoja, voidaan tehdä esimerkiksi alfanumeerinen tarkistus, eli tutkia sisältääkö kenttä myös kirjaimia. Toisaalta, jos numeerisen kentän oletetaan saavan vaikkapa vain positiivisia arvoja, voidaan tälle sarakkeelle asettaa tarkistus, joka katsoo ovatko numerot sallitun arvovälin sisällä. (McGilvray 2008, s. 124)

Validiteettia voidaan mitata myös vertaamalla datasetin jonkin tietyn kentän saamia arvoja referenssitauluun, jossa vaaditut arvot ovat saatavilla. Asiakkaan osoitetietojen validiteettia voitaisiin tarkastella esimerkiksi postinumerojen avulla, eli vertaamalla asiakastiedoissa olevia postinumeroja referenssitauluun, joka oletetusti sisältää kaikki Suomen postinumeroalueet oikein. Tämän jälkeen voidaan laskea tuolle kentälle oikeellisuusarvo jakamalla valitut rivit kokonaisrivimäärällä ja vähentämällä tämä luku halutusta maksimitarkkuusasteesta (esimerkiksi 100 prosentista).

2.1.3 Täydellisyys

Datan voidaan nähdä olevan täydellistä silloin, kun tietty datasetti sisältää kaikki tarvittavat tiedot (Batini & Scannapieco, 2016, s. 103). Täydellisyydestä (completeness) puhutaan joskus myös ”täyttöasteena” (fill-rate), eli se mittaa missä määrin kentässä ylipäätään on arvoja (McGilvray 2008, s. 296). Yksi täydellisyyden edellytys on, että tietyille sarakkeille on oltava aina arvo, eli ne eivät voi olla tyhjiä. Toinen odotus on, että datassa on kaikki asianmukaiset rivit tai sarakkeet. (Mosley et al. 2010, s. 296) Yksinkertaisimmillaan täydellisyyttä voidaan mitata laskemalla sarakkeen tyhjien (null) rivien määrä ja jakamalla se kokonaisrivimäärällä, mikä taas vähetään halutusta täydellisyysasteesta (esimerkiksi 100 prosenttia), jolloin saataisiin tuon kentän prosentuaalinen täydellisyys.

2.1.4 Ajankohtaisuus / Oikea-aikaisuus

Tietyt liiketoiminnan tarpeet ovat toisia aikakriittisempiä, jonka vuoksi näiden liiketoimintojen käyttämien datojen on oltava ajan tasalla ja saatavilla oikeaan aikaan (Mosley et al. 2010, s. 297). Tietojen arvot muuttuvat ajan myötä tai tiedot voivat olla tarpeisiin nähden vanhentuneita. Vanhentunut tieto tai se, ettei dataa ole saatavilla oikeaan aikaan voi aiheuttaa kriittisiä seurauksia, jonka vuoksi ajankohtaisuutta / oikea-aikaisuutta (currentness / timeliness) on syytä mitata. (Cai & Zhu 2015, s. 3)

Oikea-aikaisuus tai ajankohtaisuus tarkoittaa tiedon saavutettavuuden ja saatavuuden odotettua aikaa (Hassenstein & Vanella 2022, s. 6). Ajankohtaisuuden mittaamiseksi on määritettävä ajankohta, jolloin tuon datan tulee olla käytettävissä sekä on mitattava prosessin todellinen kesto alkulähteeltä käytettäväksi saakka. Eli toisin sanoen, sen mittaamiseksi on tunnistettava, kauanko kuluu aikaa siihen, että data on valmiina käytettäväksi ja verrata siihen aikaan, jolloin informaation odotetaan käytettävissä. Lisäksi on pohdittava datan volatiliteettia, joka kuvaa sitä, kuinka usein tiedon oletetaan päivittyvän. On stabiileja datalähteitä, silloin tällöin päivittyviä tietoja sekä on myös reaaliajassa päivittyvää dataa. Stabiilista tiedosta esimerkkinä toimii henkilön syntymäpäivä, kun taas reaaliaikaista tietoa voi olla vaikkapa mittauslaitteista sensorien avulla mitattu tieto. Silloin tällöin päivittyviä tietoja ovat tyypillisesti eräajot järjestelmistä, joissa reaaliaikaisuus ei ole välttämätöntä ja ajot voidaan suorittaa vaikkapa yön aikana, jolloin data on ladattuna aamuun mennessä. (McGilvray 2008, s. 32; Sebastian-Coleman 2013, s. 62, 84)

Ajankohtaisuutta voidaan esimerkiksi mitata tekemällä kohtuullisuustarkistus, eli verrataan prosessin todellista kestoa historialliseen prosessin kestoon tai määriteltyyn aikarajaan. (Sebastian-Coleman 2013, s. 87)

2.1.5 Jäljitettävyys

Datan jäljitettävyydellä (traceability) pyritään varmistamaan datan alkuperä sekä sille tehdyt transformaatiot luontihetkestä hävitykseen saakka (Zhang 2020). Siihen liittyy myös odotuksia, että data on hyvin dokumentoitu, varmistettavissa ja helposti liitettävissä alkulähteeseensä (Wang & Strong 1996, s. 28). Vaikka se ei varsinaisesti ole datalle luontainen ominaisuus, se on yksi datan laadun edellytyksistä, sillä jäljitettävyyden avulla voidaan tutkia

milloin mahdolliset virheet ovat sattuneet ja pyrkiä korjaamaan ne ajoissa. Eli datasta täytyy tietää sen alkupään lähdejärjestelmät (upstream source systems), alkuperäinen lähde (originating source), sekä kaikki loppupään järjestelmät (downstream target systems) missä dataa hyödynnetään. (Mahanti 2021, s. 126–127). Toisaalta datassa tulisi olla tarvittavat attribuutit näiden asioiden seuraamiseen, joten se ei ole myöskään täysin järjestelmäriippuvainen datan laadun ominaisuus (International Organisation of Standardization 2008, s. 17). Jäljitettävyyden määritelmän kautta, jäljitettävyyttä voidaan mitata esimerkiksi laskemalla, kuinka monesta dataputkesta datan liikkuminen alkulähteeltä loppupäähän on kuvattu sekä, että näissä dataputkissa kaikki datalle tehdyt operaatiot ovat dokumentoitu selkeästi ja muutoshistoria on tallessa. (Moses 2022, kappale 2; Mahanti 2019, s. 120–121)

2.2 Heikon datan laadun vaikutukset ja kustannukset

Tutkimukset osoittavat, että huono datan laatu laskee asiakastyytyväisyyttä sekä johtaa kasvaneisiin kuluihin ja on arvioitu, että huonosta datan laadusta aiheutuvat kustannukset voivat olla jopa 8–12 % liikevaihdosta tai palveluorganisaatiossa 40–60 % kustannuksista (Redman, 1998, s. 82). Näitä kustannuksia voi kuitenkin olla vaikea ymmärtää tai hahmottaa, mikäli laatua tai sen vaikutusta ei mitata millään tavalla. Taloudellisen analyysin tai arvion tekeminen voi itsessään olla valaisevaa mitä tulee organisaation toimintaan, tiedon laadun vaikutuksiin toimintaan ja yleiseen tietoisuuteen tiedon laadun merkityksestä organisaatiolle. Lähtökohtana on siis, että kustannusten välttämisestä tai kustannussäästöistä kertyneet hyödyt voidaan mitata, kun ne tunnistetaan. (Lee et al., 2006, s. 14–15)

Loshin (2011) esittää yksinkertaisen taksonomian, jolla voidaan arvioida heikon datan laadun negatiivisia vaikutuksia tai parantuneesta datan laadusta aiheutuvia mahdollisuuksia. Näihin lukeutuu mukaan taloudelliset, luottamukseen ja asiakastyytyväisyyteen, tuottavuuteen sekä riskiin ja vaatimustenmukaisuuteen liittyvät vaikutukset. Taloudellisilla vaikutuksilla viitataan suoraan esimerkiksi käyttökustannusten kasvamiseen, tulonmenetyksiin, tai menetettyihin mahdollisuuksiin ja kassavirran vähenemiseen. Asiakas- tai työntekijätyytyväisyyttä sekä luottamusta voi horjuttaa heikentynyt luottamus organisaatioon, esimerkiksi virheelliset ja viivästyneet päätökset. Tuottavuusvaikutuksilla viitataan muun muassa lisääntyneeseen/vähentyneeseen työmäärään tai käsittelyaikaan ja heikentyneeseen lopputuotteen laatuun. Riski ja vaatimustenmukaisuusvaikutuksiin luetaan mukaan petokset, alan

säädösten tai odotusten pettäminen, tai vaikkapa mikäli asiakastietoja on vuodettu. (Loshin, 2011, s. 6) Myös Olson (2008) nostaa esille esimerkkejä huonosta datan laadusta aiheutuvia kustannuksia ja menetettyjä mahdollisuuksia, kuten transaktioiden uudelleenkäsittelystä aiheutuneet kustannukset, uusien järjestelmien implementointikustannukset, datan ja sen avulla tehtävän päätöksenteon viivästyminen, menetetyt asiakkaat huonon palvelun vuoksi tai menetetty tuotanto toimitusketjun ongelmien vuoksi (Olson, 2008 s. 12–14).

Datan laadun kustannuksia voidaan luokitella taksonomiaan, joka on esitetty taulukossa 3. Se perustuu tyypilliseen laatukustannushierarkiaan, mikä on jaettu suoraan huonosta laadusta aiheutuviin kustannuksiin sekä laadunparannuksiin tai varmistamiseen kohdistuviin kustannuksiin. (Ge & Helfert 2013, s. 78)

Taulukko 3: Datan laadun kustannusten taksonomia (mukaillen Ge & Helfert 2013, s. 78)

Datan laadun kustannukset	Huonon datan laadun aiheuttamat kustannukset	Suorat kustannukset	Uudelleenkirjaukset
			Korvauskulut
			Tarkistuskustannukset
		Epäsuorat kustannukset	Huonon maineen aiheuttavat kustannukset
			Vääristä päätöksistä aiheutuvat kustannukset
			Uponneet investointikustannukset
	Datan laadun parantamisesta tai varmistamisesta aiheutuvat kustannukset	Ennaltaehkäisyn kustannukset	Koulutuskustannukset
			Monitorointikustannukset
			Laatustandardien kehitys ja jalkautus
		Havaitsemiskustannukset	Analysointikustannukset
			Raportointikustannukset
		Korjauskustannukset	Korjauksen suunnittelu
			Korjauksen jalkautus

Huonon datan laadun kustannukset ovat jaettu suoriin ja epäsuoriin kustannuksiin, joista suorat käsittävät esimerkiksi kustannukset, mitkä johtuvat virheellisten transaktioiden

uudelleenkirjauksesta tai ylimääräisistä tarkistuksista. Epäsuoriin kustannuksiin puolestaan lasketaan mukaan esimerkiksi puutteelliseen dataan perustuvat tehdyt päätökset tai uponneet investointikustannukset. Laadunparannuksen tai varmistamisen kustannukset ovat jaettu ennaltaehkäisyyn, havaitsemiseen sekä korjauskustannuksiin. Ennaltaehkäisyn kustannuksilla viitataan sellaisiin, joita kulutetaan välttyäkseen virheiltä tulevaisuudessa, kuten henkilöstön kouluttaminen tai laatustandardien kehittäminen. Havaitsemiskustannukset liittyvät datan laadun seurantaan ja profilointiin, joita tulee analysoinnin ja raportoinnin myötä. Arvokkaita henkilöstöresursseja voitaisiin säästää automatisoimalla toimintoja sen sijaan, että ongelmia seurataan ja analysoidaan manuaalisesti. Korjauskustannuksiin lasketaan datan laadun parannusaktiviteetit, eli ne kustannukset, joita kuluu korjauksien suunnitteluun sekä implementoimiseen. (Ge & Helfert 2013, s. 77–78)

Usein tietoja saatetaan korjata sitä mukaa kuin niitä ilmenee, mutta mikäli varsinaisiin juurisyihin ei puututa, voivat ongelmat kasaantua ja niistä aiheutua myöhemmin vielä suurempia seuraamuksia (Loshin 2011, s. 35; Mosley et al, 2010, s. 315). Huonosta laadusta aiheutuvat korjauskustannukset ovat hyvin merkittäviä verrattuna ennakoivaan laadunhallintaan käytettäviin investointeihin ja ylläpitokustannuksiin (Ge & Helfert 2013, s. 96). Ilman kustannusten mittaamista ei pystytä kuitenkaan suoraan sanomaan, saavutetaanko ennakoivalla laadunhallinnalla enemmän hyötyä kuin korjaamalla vahingot, eli on hyvä pitää mielessä näiden välinen tasapaino (Batini & Scannapieco 2016, s. 319).

3 Datan laadunhallinta

Käsiteltävän datan määrä on kasvanut merkittävästi viime vuosikymmenten aikana tietojärjestelmien ja ohjelmistojen ripeän kehityksen saralla. Järjestelmäinfrastruktuuri ja datan hallinta muuttuu entistä kompleksisemmaksi, sillä datalähteitä ja tietovirtoja on paljon. Valtava datamassa ei kuitenkaan sellaisenaan tuota käyttäjälleen varsinaisesti arvoa, vaan vasta jalostusprosessin myötä siitä tuotetaan tietoa, jota voidaan hyödyntää päätöksenteon tukena. Lisäksi huonolaatuisella datalla voidaan ajautua tilanteeseen, jossa siitä jalostetulla informaatiolla tehdään vääränalaisia päätöksiä, millä voi olla tuhoisia ja kalliita seurauksia, jonka vuoksi datan laadunhallintaa on syytä harjoittaa. (Cai & Zhu 2015)

Datan laadunhallinta (DQM) on erityisesti muutoksenhallinnan kriittinen tukiprosessi, sillä muun muassa liiketoiminnan painopisteen muuttaminen, yritysten integrointistrategiat ja fuusiot, yritysostot ja uudet kumppanuudet voivat edellyttää uusien rajapintojen ja integraatioiden luomista tai tietolähteiden yhdistämistä. DQM on myös jatkuva, kokonaisvaltainen ja organisaatiota poikkileikkaava toiminto, jonka puitteissa määritetään hyväksyttävä datan laadun taso liiketoiminnan tarpeisiin nähden ja varmistetaan, että todellisuudessakin laatu vastaa näitä tasoja. DQM sisältää datan laadun analysoinnin, tietojen poikkeamien tunnistamisen sekä liiketoimintavaatimusten ja vastaavien liiketoimintasääntöjen määrittelyn vaaditun datan laadun takaamiseksi. Se pitää siis sisällään muitakin toimenpiteitä, kuin datan korjaamisen tai valvonnan. (Mosley et al. 2010, s. 291)

Onnistunut datan laadunhallinta vaatii liiketoiminnan panoksen siitä, mitkä asiat ovat kriittisiä ja mitä asioita on syytä seurata, mutta myös IT:n panoksen mahdollistaa datan laadun mittaaminen, monitorointi sekä käynnistää datan jäsentäminen, standardointi, puhdistaminen ja konsolidointi tarvittaessa. Näiden lisäksi tärkeää on luoda DQM:n eri rajapintojen tai sidosryhmien välille yhteisymmärrys esimerkiksi siitä, millä kanavilla laatuun liittyvistä asioista kommunikoidaan ja prosessit siihen, miten ne ratkotaan tai niiden kanssa menetellään. (McGilvray 2013, s. 43)

3.1 Datan laadunhallinnan viitekehyksiä ja menetelmiä

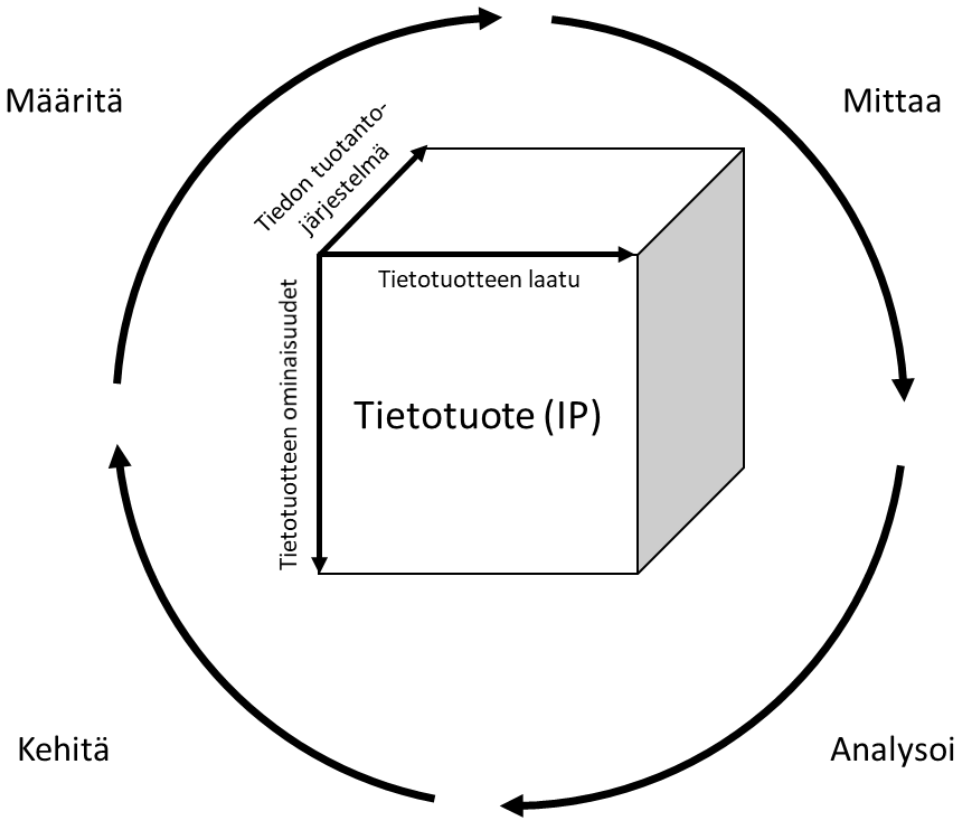
Kirjallisuudessa esitetään useita DQM viitekehyksiä ja hallintamalleja (taulukko 4), joissa on samanlaisia piirteitä, mutta myös eroavaisuuksia esimerkiksi sen suhteen, kuinka niitä hyödynnetään. Tyypillisesti kuitenkin datan laadun viitekehykset sisältävät kuvauksen datan laadun dimensioista sekä siitä, miten datan laatua arvioidaan ja kuinka sitä voidaan parantaa. Osa viitekehyksistä ovat geneerisiä, eli sellaisenaan hyödynnettäviä, toiset niin sanotusti modulaarisia, eli niitä täytyy soveltaa tilanteen ja organisaation tarpeiden mukaan. Lisäksi osa viitekehyksistä ja tutkimuksista ovat teoreettisia, eli niiden soveltuvuutta käytännössä erilaisissa organisaatioissa ei ole testattu tai niiden toimivuutta ei tiedetä. Jotkut niistä, kuten Data Management Book of Knowledge (DAMA DMBOK), on pääsääntöisesti suunnattu datan hallintaan, mutta sisältää erillisen datan laadunhallinnan osion.

Taulukko 4: Datan laadunhallinnan viitekehyksiä/menetelmiä

Viitekehys/menetelmä	Lähde	Pääkomponentit
Total Data Quality Management (TDQM)	Wang 1998	TDQM-sykli: määritä, mittaa, analysoi, kehitä.
Total Information Quality Management (TIQM)	English 1999	Luo edellytykset datan laadun mittaamiseksi, arvioi datan definitiot ja arkkitehtuurin laatu, mittaa datan laatu sekä ei-laadulliset datakustannukset, uudelleensuunnittele ja korjaa data, paranna dataprosessien laatua.
The Ten Steps	McGilvray 2008	Kymmenenvaiheinen prosessi: määritä liiketoimintatarpeet ja lähestymistapa, analysoi tietoympäristö, arvioi datan laatu sekä vaikutus liiketoimintaan, tunnista juurisyyt, luo kehittymissuunnitelma, ennaltaehkäise datavirheet, korjaa nykyiset datavirheet, jalkauta kontrollit.
DAMA DMBOK (Data Quality Management-osio)	Mosley et al. 2010	DQM-sykli: suunnittelu (planning), kontrolli (control), kehittäminen (development), operatiivinen toiminta (operational). Kaksitoista aktiviteettia syklin vaiheiden alla.
Data Quality Assessment Framework (DQAF)	Sebastian-Coleman 2013	Kerran toteutettava arviointi, automatisoidut prosessikontrollit, automaattinen prosessin aikana tehtävä mittaus (in-line), jaksoittain tapahtuva mittaus.
Framework for Data Quality Management (based on ISO 8000-61)	King & Schwarzenbach 2020	ISO 8000-61-standardiin pohjautuva viitekehys, joka sisältää seuraavat vaiheet: datan laadun suunnittelu-, kontrolli-, varmistus- ja parannus.

Merkittävimmät datan laadusta tehdyt tutkimukset ulottuvat 1990-luvun alkupuolelle, joihin aikoihin Massachusettsin teknillinen instituutti (MIT) käynnisti kokonaisvaltaisen datan laadunhallinnan (TDQM) -ohjelman. Ohjelman tarkoituksena oli korostaa datan laatuun keskittyvän tutkimuksen merkitystä ja se toimiikin yhä perustana nykytutkimuksille. (Madnick et al. 2009, s. 2) Kokonaisvaltaisessa datan laadunhallinnassa dataa tai informaatiota lähestytään tuotenäkökulmasta, eli keskitytään hallitsemaan koko datan elinkaarta ja sen aikana

tapahtuvia aktiviteetteja, kuten datan syntymistä, sen muokkaamista ja liikkumista, joista jokaisesta on pidettävä huolta, jotta datan laatu vastaa sen käyttäjien odotuksia (Mosley et al. 2010, s. 291). TDQM metodologia perustuu iteratiiviseen prosessiin, jonka peruspilarit ovat Lean Six Sigmasta lainatut DMAIC-viitekehyksen komponentit, määritä (define), mittaa (measure), analysoi (analyze), kehitä (improve) sekä ohjaa (control). Metodologia on esitetty kuvassa 3. Datan laadunhallinnassa saatetaan toisinaan myös hyödyntää Deming-sykliä, jonka vaiheet ovat suunnittele (plan), tee (do), tarkista (check), tee muutokset (act), kuten esimerkiksi mihin DAMA DMBOK DQM-sykli pohjautuu (Mosley et al. 2010, s. 292–293).

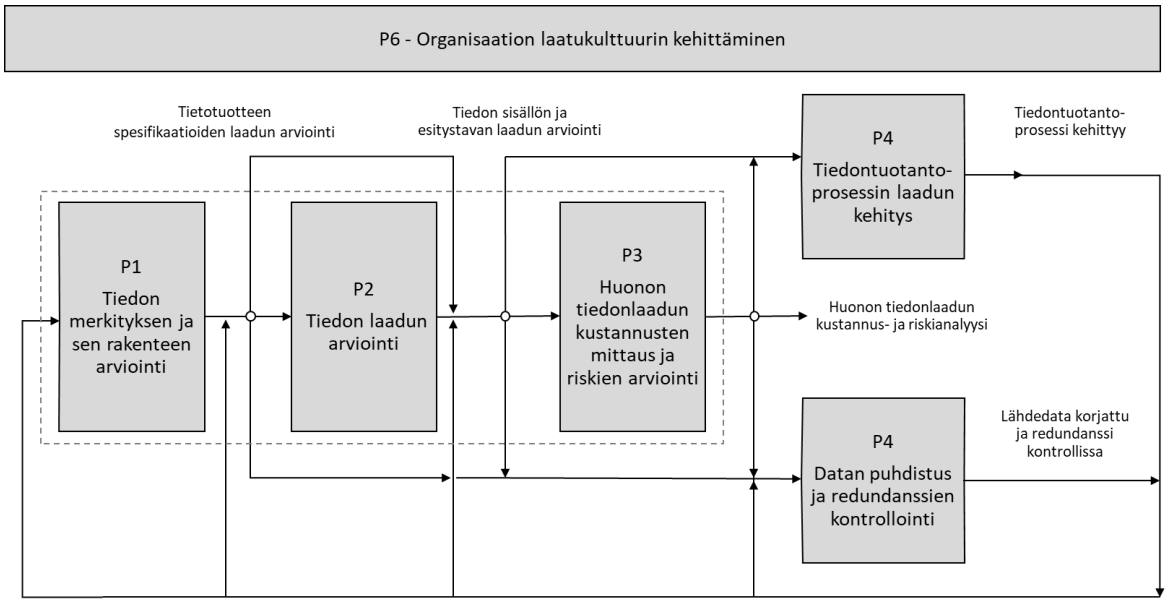


Kuva 3: TDQM metodologia (mukaillen Wang 1998)

Dataa syntyy tuotoksena liiketoimintaprosesseista, joita IT-yksikkö ei omista, joten liiketoimintojen täytyy pitää huolta datan syöttämisen kannalta olennaisista laadunhallinnan korjausprosesseista ja ohjeistuksista (Redman 2013, s. 16, 26). Liiketoiminnan on oltava mukana syklin ensimmäisessä vaiheessa, eli määrittämässä mitattavia asioita, sillä heillä on

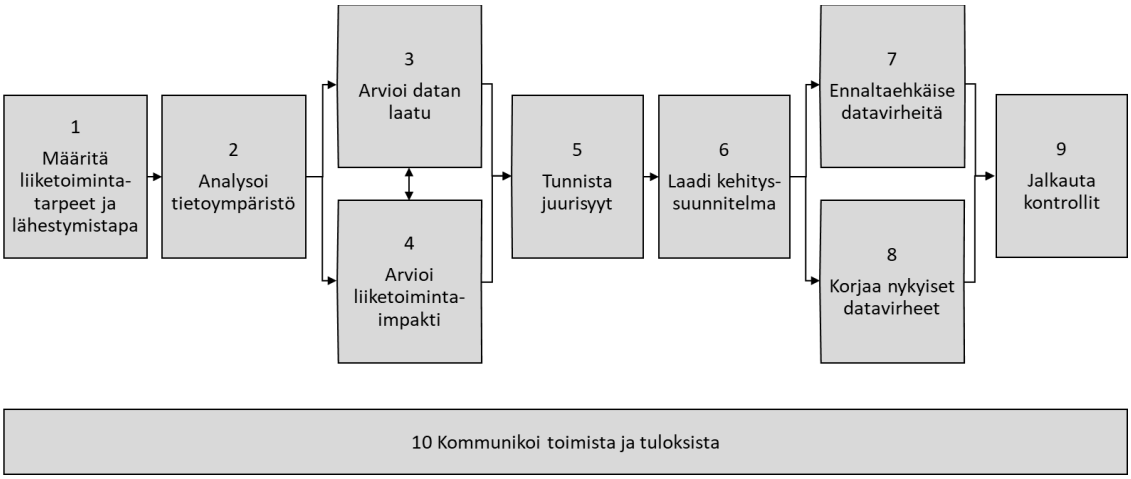
eniten ymmärrystä datan kontekstista. IT:n tehtävänä on pystyttää tarvittavat teknologiat, jotta laatua voidaan mitata seuraavassa vaiheessa. Tietojen analysoinnilla pyritään saamaan kiinni virheiden juurisyistä ja viimeisessä vaiheessa pyritään kehittämään toimintatapoja perustuen löydettyihin syihin. TDQM-metodologiassa tunnistetaan myös neljä olennaista roolia, jotka ovat tiedon toimittajat (suppliers), tiedontuottajat (manufacturers), tiedon kuluttajat (consumers) sekä tietotuotteesta vastaavat managerit.

Kokonaisvaltainen tiedon laadunhallinta (TIQM) pitää sisällään enemmän kuin pelkän datan laadun parannuksen tai sen korjaamisen prosessit, sillä se nähdään kaikkien organisaatiossa olevien vastuulla olevana jatkuvan kehityksen prosessina. TIQM:ssä on kuusi pääprosessia (kuvassa 4), jotka ovat: P1 - tiedon merkityksen ja sen rakenteen arviointi, P2 - tiedon laadun mittaaminen ja arviointi, P3 - huonon tiedonlaadun kustannusten mittaus ja riskien arviointi, P4 - tiedontuotantoprosessien laadun kehittäminen, P5 - datan korjaus ja redundanssien kontrollointi, sekä P6 - laatukulttuurin vakiinnuttaminen organisaatiossa. (English 2009; Francisco et al. 2017)



Kuva 4: TIQM-viitekehys (Mukaillen English 2009)

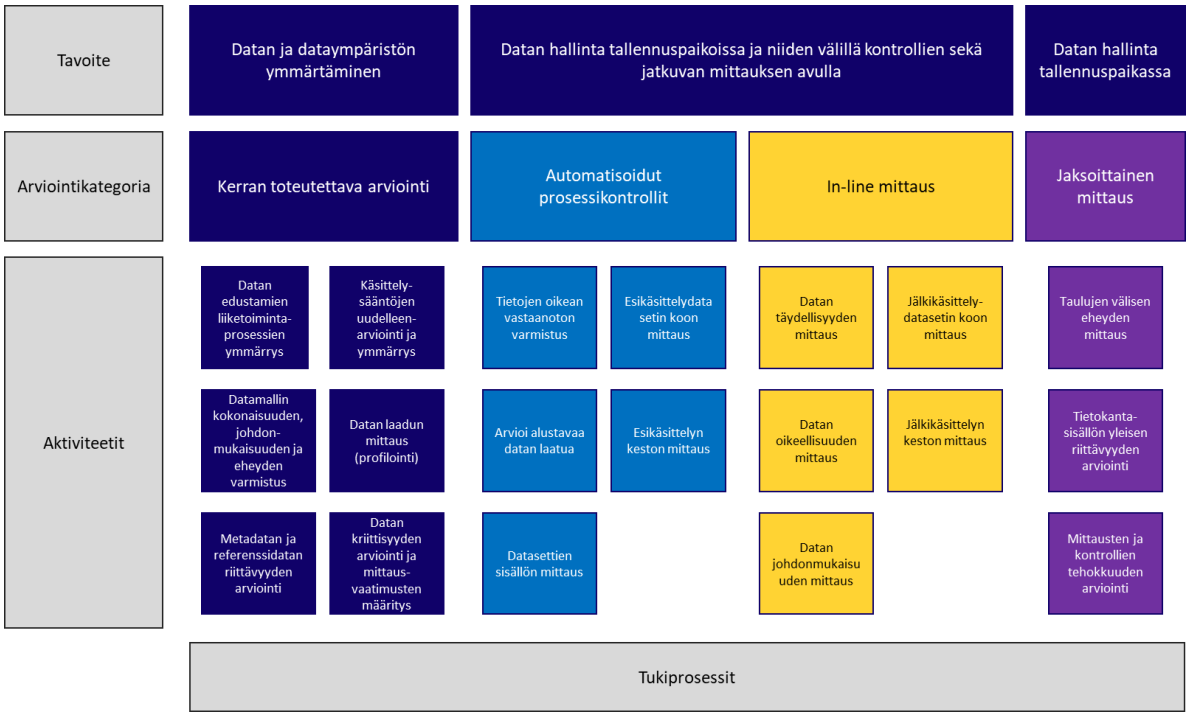
McGilvray (2008) esittää kymmenenvaiheisen datan laadunhallinnan viitekehyksen (kuvassa 5), jota organisaatiot voivat hyödyntää arvioimaan, parantamaan ja luomaan tietoa ja datan laatua. Kirja sisältää jokaisen vaiheen osalta konkreettisia ohjeita, kuinka suunnitellaan ja implementoidaan tiedon tai datan laadun parannusprojekteja.



Kuva 5: The Ten Steps process, kymmenenvaiheinen datan laadunhallinnan prosessi. (Mukaillen McGilvray 2008, s. 64)

Prosessi lähtee liikkeelle liiketoimintatarpeiden kartoituksella ja lähestymistavan valinnalla, eli määritellään ja päätetään ongelma, mahdollisuus tai tavoite, joka ohjaa toimintaa projektin aikana. Toisessa vaiheessa kootaan ja analysoidaan kaikki tiedot nykyisestä tilanteesta ja tietoympäristöstä, jotta voidaan luoda hyvä pohja seuraaville prosessin vaiheille ja mikä varmistaa, että käsitellään tilanteeseen sopivia ja tarpeenmukaisia tietoja. Kolmas vaihe pitää sisällään varsinaisen datan laadun arvioinnin, eli tarkistetaan määritettyjen datan laadun dimensioiden avulla millä tasolla datan laatu on. Myöhemmissä vaiheissa tuloksia voidaan hyödyntää muun muassa juurisyiden ja parannuskohteiden tunnistamiseksi. Seuraavaksi prosessin neljännessä vaiheessa arvioidaan huonon datan vaikutus liiketoimintaan. Tämän vaiheen tuloksilla voidaan muun muassa selvittää liiketoiminnan kannattavuuden parantamisen kannalta olennaiset asiat ja siten saada tukea ja määrittää asianmukaiset resurssit laadunparannushankkeille. Viidennessä vaiheessa tunnistetaan ja priorisoidaan todelliset juurisyyt dataongelmille ja luodaan spesifit suositukset, kuinka niiden osalta toimitaan. Kuudennen vaiheen aikana laaditaan kehityssuunnitelma, eli viimeistellään aiemmassa vaiheessa luodut suositukset ja päätetään kehitystoimet perustuen näihin suosituksiin. Seitsemännessä vaiheessa otetaan käyttöön dataongelmien ennaltaehkäisyyn tarkoitetut ratkaisut, jonka jälkeen kahdeksas vaihe ottaa kantaa nykyisten ongelmien korjaukseen. Implementoitujen parannusten monitorointi ja verifiointi tapahtuu yhdeksännessä vaiheessa, jonka jälkeen lopuksi dokumentoidaan ja kommunikoidaan laatutestien tulokset, tehdyt parannukset ja parannuksien lopputulokset. Kommunikointi ei kuitenkaan ole ainoastaan viimeisen vaiheen toimi, vaan sitä tulee tehdä prosessin jokaisessa vaiheessa.

Sebastian-Colemanin (2013, s. 57) DQAF-viitekehyksen tarkoituksena on auttaa organisaatioita arvioimaan ja parantamaan tiedon sekä datan laatua mittareiden avulla. DQAF-viitekehys on esitetty kuvassa 6. Se sisältää neljä arviointikategoriaa, jotka ovat: kerran toteutettava arviointi, automatisoidut prosessikontrollit, tiedonkäsittelyprosessin aikainen jatkuva mittaus (in-line measurement) ja jaksoittain tapahtuva mittaus.

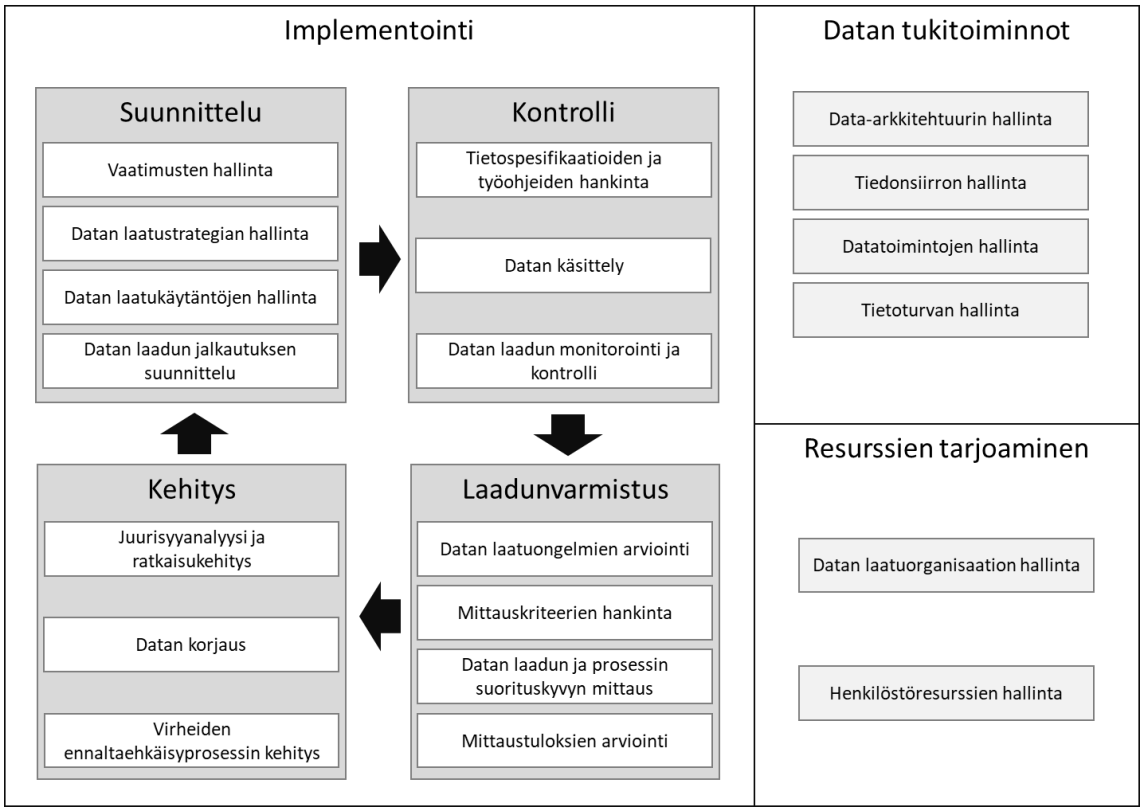


Kuva 6: DQAF-viitekehys (mukaillen Sebastian-Coleman 2013, s. 67)

Kertaluontoisen arvioinnin tavoitteena on auttaa organisaatiota hahmottamaan käyttämäänsä dataa ja ympäristöä, missä sitä käytetään. Tyypillisesti tässä vaiheessa syvennytään muun muassa profilointiin, jonka avulla datan sisällöstä, struktuurista ja kunnosta saadaan hyödyllistä metadataa. Tämän vaiheen tuloksien avulla tuotetaan ehdotuksia laadunparannuksen osalta. Automatisoidun prosessikontrollin tarkoitus on pitää järjestelmä vakaana. Sen avulla saadaan palautetta mahdollisista epävakauksista, joihin perustuen se panee aluilleen jonkin määrätyn toimen. Tiedonsiirron rivimäärien varmistaminen (lähetetyt vs. kirjoitetut) tai latauksen kesto voi olla yksi esimerkki tällaisesta tarkistuksesta. Mikäli virheitä tai epävakauksia havaitaan, niihin perustuva määrätty toimi olla vaikkapa virheilmoituksen sisältävän sähköpostin lähettäminen automaattisesti datavastaavalle. Jatkuvan mittauksen kohteiksi kannattaa valita liiketoimintakriittiset tai usein päivittyvät tiedot, jotta mahdollisiin virheisiin voidaan varautua nopeammin. Jaksoittain tehtävät tarkistukset ovat usein hyvä toteuttaa ei-kriittisille datalähteille, jotka eivät vaadi jatkuvaa mittausta. Vertauskohtana voidaan

hyödyntää profiloinnin tai aiempien mittausten tuloksia, jotta voidaan tutkia, onko data muuttunut viime kerrasta. (Sebastian-Coleman 2013, s. 124–125)

ISO 8000-61 standardi (esitetty kuvassa 7) kuvaa kokonaisvaltaista lähestymistapaa datan laadunhallintaan. Sen mukaan datan laadunhallinta pitää sisällään 20 teknologiasta riippumatonta alaprosessia, jotka myös ottavat huomioon menettelytavat sekä vastuullisuuden, eli kuinka varmistetaan onnistunut datanhallinta (data governance).



Kuva 7: ISO 8000-61 prosessimalli (mukaillen King & Schwarzenbach 2020, s. 58).

Datan laadun suunnitteluvaiheessa määritetään yleiset vaatimukset, tavoitteet ja suunnitelma kuinka päästään tavoiteltuun maturiteettitasoon datan laadunhallinnan osalta organisaatiossa. Tässä vaiheessa voidaan laatia esimerkiksi datan laadun strategia, erilaisia menettelytapoja sekä jalkautusvaiheen suunnitelma. Kontrollivaiheessa otetaan huomioon kaikki ne aktiviteetit, joilla voidaan määrittää ja kontrolloida datan syntyä ja päivitystä, että laatu vastaa haluttuja vaatimuksia. Kontrollivaiheessa pyritään kartoittamaan, että prosessien tietovaatimukset ovat määritetty esimerkiksi tutkimalla tietospesifikaatioita ja työohjeita. Datan laadun seuranta luetaan myös tämän prosessin aktiviteetiksi, eli toisin sanoen ne toimet, joilla datan laatua seurataan ja valvotaan, jotta voidaan tunnistaa ja reagoida tapauksiin, joissa käsittely ei ole ollut vaatimustenmukaista. Datan laadunvarmistus kattaa ne toimet,

joilla pyritään ymmärtämään datan laadun nykytilaa, kuten tyypillisten virheiden tunnistaminen ja arviointi, laadun mittarien ja mittaamista tukevien metodien määritys, datan laadun mittaaminen ja prosessien suorituskyky sekä mittaustulosten arviointi. Datan laadun kehitysvaiheessa käynnistetään datan laadunparannusaktiviteetit, kuten juurisyyanalyysi ja ongelmanratkaisu, datan korjaus, sekä prosessikehitys tavoitteenaan ennaltaehkäistä virheitä. Näiden prosessien lisäksi on erilaisia datan hallinnan tukiprosesseja tai aktiviteetteja, kuten data-arkkitehtuurin, tiedonsiirron, datan operoinnin sekä datan tietoturvan hallinta. Organisaation kyvykkyys on myös olennainen osa ISO 8000-61 standardia, eli kuinka datan hallinnan ja dataan liittyvän kehittämisen taitoja kehitetään, jotta pystytään tarjoamaan resursseja organisaatiotasolla. Osaamista voidaan kehittää esimerkiksi muodostamalla datan laatuorganisaatioita, eli jakaa parhaita laadunhallintakäytäntöjä ja kehittää taitoja organisaatiolaajuisesti esimerkiksi organisaatioyksikköjen, komiteoiden ja avainhenkilöstöverkostojen kautta. (King & Schwarzenbach 2020, s. 57–59)

3.2 Datan laadunhallinnan pääperiaatteet

Aiemmin esitellyissä DQ-viitekehyksissä toistuu samankaltaisia piirteitä ja aktiviteetteja, joita tulisi ottaa huomioon datan laadunhallintaa suunniteltaessa ja implementoitaessa. Nämä voidaan karkeasti jakaa laatukirjallisuudestakin tuttuun viiteen pääprosessiin, jotka ovat määritä, mittaa, analysoi, kehitä ja ohjaa. Näiden prosessien alle taas voidaan asettaa erilaisia aktiviteetteja, kuten liiketoimintavaatimusten kartoittaminen kyselyillä tai haastatteluilla, roolien ja vastuiden määrittäminen, datan profilointi ja muiden tarvittavien työkalujen käyttöönotto tai teknologioiden pystyttäminen, datan laadun mittaaminen sekä tulosten visualisointi ja seuranta (Loshin 2011, s. 20). Näiden lisäksi on oltava foorumi, missä dataongelmista keskustellaan ja päätetään, miten niiden kanssa toimitaan. Toimintoja pitää pystyä toteuttamaan mahdollisimman läpinäkyvästi, suoraviivaisesti sekä yhteisesti määritettyjen toimintatapojen mukaisesti, jotta laadunhallinnasta saatava hyöty voidaan maksimoida. (Loshin 2011, s. 36-37)

3.2.1 Määritä

Määritysvaiheessa datan laadusta ensisijaisesti vastuussa oleva tiimi arvioi tunnettujen ongelmien laajuutta, mikä sisältää ongelmien kustannusten ja vaikutusten määrittämisen sekä vaihtoehtojen arvioinnin niiden ratkaisemiseksi. Määritys alkaa tunnistamalla liiketoimintatavoitteiden saavuttamisen tiellä olevat kriittiset dataongelmat, määrittelemällä liiketoiminnan vaatimukset datan laadulle, tunnistamalla tärkeimmät datan laadun ulottuvuudet ja määrittelemällä liiketoimintasäännöt, jotka ovat kriittisiä korkean laadun varmistamiseksi. (Mosley et al. 2010, s. 293) Tässä vaiheessa voidaan arvioida myös dataspesifikaatioiden laatua, eli tutkia, kuinka selkeästi datan käsitteet ovat mallinnettu, onko tietomallit kuvattu, liiketoimintasäännöt määritetty, tai esimerkiksi metadatan kuvaukset olemassa (McGilvray 2008, s. 117; English 2009, s. 8; King & Schwarzenbach 2020, s. 71). McGilvray (2008, s. 29) esittää kymmenenvaiheisessa prosessissaan myös vastaavassa vaiheessa neljän näkökulman (data, prosessit, ihmiset ja organisaatio, teknologia) ja kysymyksien (mitä, miten, kuka, kuinka) avulla mitä asioita tulisi ottaa huomioon, kuten mitkä ovat liiketoimintatavoitteet, mitä dataa liiketoiminta tarvitsee, mitkä ovat datastandardit, kuka kehittää prosesseja ja liiketoimintasääntöjä, mikä on korkean tason arkkitehtuuri ja mikä teknologia tukee liiketoimintaa.

Sebastian-Coleman (2013, s. 94) korostaa, että laadunparannusta edeltää arviointivaihe, jonka tarkoituksena on rakentaa kattava ymmärrys organisaation datapääomasta, tunnistaa kriittiset tai korkean riskin data, sekä jatkuvaa seurantaa vaativa data. Näiden pohjalta saadaan kirkastettua datan laadun edellytykset ja dokumentoitua nykytila sekä parannusehdotukset. Myös Mahanti (2019, s. 323) painottaa, että yrityksen tulisi keskittyä dataongelmiin, jotka kohdistuvat liiketoimintakriittiseen dataan. Esimerkiksi vähittäiskauppatoimijalla tuotetietojen laatu on kriittistä, sillä ajantasainen ja tarkka tieto tämänhetkisestä varastotilanteesta on olennainen osa kustannusten ja varastojen hallintaa sekä myyntiä.

3.2.2 Mittaa

Mittausvaiheessa valitaan ensisijaiset datan laadun ulottuvuudet, joita mitattavan asian osalta aletaan seurata (Lee et al. 2009, s. 29). Olennaista on hyödyntää määritysvaiheen tuloksia ja valita tärkeimmät datan laadun ulottuvuudet, joiden seuraamisesta oletetaan

tuottavan eniten arvoa. Profilointi toimii mittausvaiheen alussa erinomaisena työkaluna, sillä sen avulla saadaan todellista tietoa datan sisällöstä, eli siitä miten arvot jakautuvat, onko tiedoissa puuttuvia arvoja, saavatko tietyt kentät tyhjiä (null) arvoja, mitkä kentät ovat uniikkeja ja niin edelleen (Mahanti 2019, s. 325). Näitä tietoja voidaan käyttää avuksi määrittämään raja-arvoja datan laadun mittareille sekä kvantifioida liiketoimintasäännöt, eli toisin sanoen hahmottaa miten konkreettisesti mitataan haluttuja asioita. (Sebastian-Coleman 2013, s. 49) Profilointia voidaan ja kannattaakin suorittaa aika ajoin, sillä data saattaa muuttua ja liiketoimintasäännöt sen mukana. Ensimmäisen perusprofiloinnin jälkeen valitaan tyypillisesti säännölliseen seurantaan ne tiedon laatuongelmat, joilla on suurin vaikutus (McGilvray 2008, s. 120).

Kerroksittain rakennetussa ja useita käsittelykerroksia sisältävässä data-arkkitehtuurissa datan laatua olisi hyvä seurata jokaisella tasolla. Eri kerroksilla on usein omat tarkoituksensa, kuten duplikaattien poisto, datatyyppimuunnokset, liiketoimintalogiikan lisääminen, useiden tietolähteiden yhdistäminen toisiinsa yhtenäiseksi tietomalliksi, joista syistä erilaisia mittareita kerroksittain on hyvä olla. Lisäksi läpi käsittelyprosessin täytyy seurata niin sanottuja dataputkien terveysmittareita, eli esimerkiksi laskea kuinka monta riviä on luettu vs. kirjoitettu, onko kaikki data saatu luettua (volyymi), tai kauanko ajot kestävät ja milloin ne ovat viimeksi ajettu (tuoreus). (Moses 2022, kappale 2)

Mittausvaiheessa tulisi kartoittaa ja visualisoida myös datan kulku ja sille tehdyt toimenpiteet alkulähteestä loppupäähän. Tällainen datan elinkaaren visualisointi (data lineage) helpottaa myös virheiden selvitystyötä erityisesti monimutkaisissa arkkitehtuureissa. Datan elinkaaren visualisointi on yksi havaittavuuden (observability) viidestä pilarista, joilla voidaan tutkia dataputkien kuntoa. Muita pilareita ovat tuoreus (freshness), jakauma (distribution), volyymi (volume), sekä skeema (schema). Jakauma kertoo, onko data sallittujen arvojen sisällä tai oikeassa muodossa. Skeema viittaa nimensä mukaisesti siihen, onko sen rakenne muuttunut, eli onko dataan tullut tai onko siitä poistunut sarakkeita. (Moses 2022, kappale 2)

3.2.3 Analysoi

Analysointivaiheessa seurataan mittauksen tuottamaa statistiikkaa datan laadusta määritettyjen mittarien puitteissa ja verrataan määritettyihin raja-arvoihin tai sääntöihin (Sebastian-

Coleman 2013, s. 110; Treder 2020, s.187). Tuloksista pyritään löytämään korrelaatioita ja toistuvia datavirheitä, joiden impakti liiketoiminnalle, juurisyyt ja ratkaisuohjeet tunnistetaan ja dokumentoidaan (McGilvray 2008, s. 109).

Datan laadun mittaustuloksien avulla voidaan tehdä johtopäätöksiä virheiden synnystä. Tärkeintä on tunnistaa juurisyyt ja lähtökohtaisesti pyrkiä korjaamaan datan laatu lähdejärjestelmässä tai vähintäänkin mahdollisimman aikaisessa vaiheessa datan elinkaarta (Mosley et al. 2010, s. 293; Moses 2022, kappale 4). Juurisyyanalyysissa (RCA) olisi syytä seurata johdonmukaista prosessia, jotta ongelmat voidaan havaita mahdollisimman pian ja tyypilliset virhetilanteet sekä ohje niiden ratkomiseksi voidaan dokumentoida. (Moses 2022, kappale 4) RCA:ssa voidaan käyttää taulukossa 5 esitettyjä metodeja, esimerkiksi ”Five Whys” metodologiaa tai kalanruotokaaviota.

Taulukko 5: Juurisyyanalyysin metodeja (mukaillen McGilvray 2008, s. 200)

Nro	Nimi	Kuvaus
1	Five Whys	Kysytään viisi kertaa ”miksi?”, jotta päästään perimmäiseen syyhyn kiinni.
2	Track and Trace	Tunnistetaan ongelman sijainti jäljittämällä tietojen elinkaarta ja määrittämällä perimmäiset syyt, joissa ongelma ilmenee ensimmäisen kerran.
3	Kalanruotokaavio	Tunnistetaan, tutkitaan ja visualisoidaan graafisesti ongelman juurisyyt omiksi ”kalanruodoiksi”. Jokaiseen syyhyn kirjataan siihen liittyviä ongelmia, jotka saadaan selville kysymällä, miksi tämä ongelma syntyy ja mikä siihen vaikuttaa.

Five Whys-metodilla pyritään viiden miksi-kysymyksen avulla pääsemään kiinni juurisyyhyn. Esimerkiksi jos havaitaan, että asiakasmasterdatassa on duplikaatteja, saatetaan ongelmaa lähestyä seuraavien viiden kysymyksen avulla (McGilvray 2008, s. 201–202):

- ”miksi asiakasdatassa esiintyy tuplarivejä?” -> koska asiakaspalvelijat luovat uusia rivejä vanhojen päivittämisen sijaan.

- ”miksi he luovat uusia rivejä vanhojen päivittämisen sijaan?” -> koska he eivät halua etsiä olemassa olevia tietoja.
- ”mikseivät he halua etsiä niitä?” -> koska niiden etsimisessä kestää liian kauan.
- ”miksi etsimisessä kestää kauan?” -> asiakaspalvelijoita ei ole koulutettu hyödyntämään hakutoimintoja tehokkaasti.
- ”miksi se on ongelma, että tietojen hakemisessa kestää kauan?” -> asiakaspalvelijoita mitataan sillä perusteella, kuinka nopeasti se saavat tiedot kirjattua järjestelmään / saavat transaktion valmiiksi, jolloin datan laatua ei palkita, eikä myöskään heillä ole näkyvyyttä siihen, miksi duplikaattitiedot olisivat ongelma muualla organisaatiossa.

Esimerkki tyypillisestä dataongelmasta on, kun data on myöhässä tai se ei ole päivittynyt ollenkaan. Tällainen tapaus voi johtua useista eri syistä, kuten transformaatiologiikan muuttumisesta, skeemamuutoksista, ajonaikaisesta virheestä, tai tietojen syöttövirheestä / vääränlainen syöttölogiikka (Sebastian-Coleman 2013, s. 115). Mitä monimutkaisempi arkkitehtuuri tai kompleksisempi tietomalli, sitä haastavampaa voi olla löytää ongelman aiheuttava juurisyy. Datatiimi voi lähteä selvittämään ongelmaa vaikkapa katsomalla datan elinkaarta ja tutkia kerros kerrokselta, missä ongelma on syntynyt (Loshin 2011, s. 216). Rivimäärämuutokset tai tunnuslukujen tahaton muuttuminen eri kerrosten välillä ovat hyvä indikaattori virheen sijainnista. Jahka kerros on selvillä, voidaan tutkia esimerkiksi dataputkien transformointilogiikkaa ja onko sen koodiin tehty muutoksia, mitkä ovat voineet ongelman aiheuttaa. Näiden lisäksi tulisi katsastaa toimintaympäristö tai dataputkien orkestrointityökalun tila sekä sieltä saadut ajojen lokitiedot ja esimerkiksi pilviympäristön palveluntarjoajan palvelinten status. (Moses 2022, kappale 4; Loshin 2011, s. 215–216) Yleisesti ottaen esittämällä seuraavat kysymykset voi päästä ongelman juureen kiinni:

- Missä virhe esiintyy ensimmäisen kerran? Seuraa elinkaaren visualisaatiota, jotta virheen ensiesiintymisen sijainti saadaan selville.
- Onko ympäristössä tapahtunut muutoksia, jotka olisivat voineet aiheuttaa virheitä järjestelmään?
- Onko prosessissa ongelmia, jotka olisivat voineet aiheuttaa häiriön?
- Onko ulkoisilla palveluntarjoajilla / toimijoilla ongelmia, jotka ovat vaikuttaneet datan laatuun?

Ongelman selvityksen toimintamalli tulee olla yksityiskohtaisesti dokumentoitu ja sitä tulisi pyrkiä aina noudattamaan systemaattisesti. Selvitysprosessi voi olla esimerkiksi seuraavanlainen (Loshin 2011, s. 239):

1) ongelma havaitaan
2) ilmoita sidosryhmille ongelmasta ja että sitä tutkitaan
3) kirjaa ongelma ylös sovittuun paikkaan
4) diagnosoi ongelman syy (RCA)
5) arvioi vaihtoehdot ongelmanratkaisuun
6) korjaa data
7) kehitä prosessia eliminoimaan ongelmat jatkossa

3.2.4 Kehitä

Kehittämisvaiheessa pyritään rakentamaan ja implementoimaan laadunparannuksen ratkaisut. Ratkaisut voivat olla teknisiä toteutuksia, henkilöstön kouluttamista, prosessimuutoksia, tai toimintamallien muutoksia esimerkiksi datan syöttöohjeisiin (Mahanti 2019, s. 326). Tärkeää on, että ratkaisut keskittyvät korjaamaan tunnistettuja juurisyitä (McGilvray 2008, s. 213). Toisinaan juurisyitä ei kuitenkaan ole kannattavaa tai mahdollista korjata, jolloin datan korjaaminen / imputoiminen tai vain virheistä hälyttäminen ovat vaihtoehtoisia toimintatapoja (Mosley et al. 2010, s. 305).

Selvitystyössä voidaan hyödyntää tapaustenhallintaa (incident management), mikä voi auttaa ongelmien arviointia, nopeuttaa alustavan diagnosoinnin tekemistä, sekä parantaa kommunikointia eri tahojen välillä (Mosley et al. 2010, s. 303). Ongelman havaitseminen ja siitä oikeille sidosryhmille tiedottaminen tulisi automatisoida mahdollisimman pitkälle, sillä se nopeuttaa ongelmanratkaisuprosessia. Tämä voi tapahtua esimerkiksi niin, että datan laadun sääntörikkomus laukaisee hälytyksen, jonka perusteella luodaan tiketti tapaustenhallintatyökaluun. (Moses 2022, kappale 4) Lisäksi näin tallentuu tieto siitä, kuka on vastuussa mistäkin tiketistä ja paljonko selvitykseen on kulunut resursseja, eli prosessin läpinäkyvyyskin lisääntyy. Seuranta mahdollistaa prosessin suorituskyvyn seurannan, mukaan lukien ongelmien

ratkaisemiseen kuluva aika, ongelmien esiintymistiheys, ongelmatyypit, ongelmien lähteet ja yleiset ongelmien korjaamisen lähestymistavat. (Mosley et al. 2010, s.303)

Näiden tietojen avulla voidaan myös kerätä historiatietoa usein toistuvista ongelmista, sekä dokumentoida yleiset ongelmat ja kuinka ne selvitettiin, mikä nopeuttaa selvitystyötä tulevaisuudessa (Loshin 2011, s. 232). Yleisistä ongelmista tulee tiedottaa kaikkia prosessiin kuuluvia ja tiedonkäsittelystä vastaavia, jotta voidaan yhdessä kehittää prosessia niin, että tyypillisistä ongelmista päästään eroon. (Mosley et al. 2010, s.307) Riippuen ongelman vakavuudesta tai laajuudesta, saattaa myös organisaatiotasolla tiedottaminen olla paikallaan (McGilvray 2008, s. 217).

3.2.5 Ohjaa

Ohjausvaiheessa tarkoituksena on jatkuvasti kehittää prosessia ja seurata datan laadun kehitystä, ajatuksena, että datan laadun ongelmiin päästään kiinni mahdollisimman aikaisessa vaiheessa. Tärkeää on myös, että resurssit keskitetään liiketoimintakriittisen datan laadun seurantaan. (Mahanti 2019, s. 325; 345) Laadun seurantakyvykkyydet antavat loppukäyttäjille luottamusta päätöksenteossa hyödynnettyyn dataan, ja siihen, että päätökset itsessään tuottavat asianmukaisia liiketoimintatuloksia. Mahdollisimman pitkälle automatisoitu monitorointi myös helpottaa suuren datamassan laadunvalvontaa ilman manuaalista työtä, vaikkakin jonkinlaista käsin tehtävää tarkistusta on hyvä edelleen olla, jotta voidaan olla varmoja siitä, että tarkistukset toimivat. (King & Schwarzenbach 2020, s. 73–74)

Ohjauksen tueksi voidaan tuottaa raportointia, mistä laatuvastuussa olevat henkilöt voivat seurata datan laadun kehitystä (Mosley et al. 2010, s. 307, s. 312). Kuten sanottu, monitoroinnin tueksi on hyvä rakentaa mahdollisimman pitkälle automatisoitu hälytysjärjestelmä, joka havaitsee suurimman osan ongelmista automaattisesti sekä hälyttää tarvittavia tahoja ongelmista. Näin aikaa ei kulu tarpeettomaan manuaaliseen seurantaan. (Sebastian-Coleman 2013, s. 124–125; Moses 2022, kappale 4; Loshin 2008, s. 308) Tulokset tulisi olla esitettynä helposti saavutettavasti, selkeästi ja läpinäkyvästi, kuitenkin niin, ettei tarpeettomilla tahoilla ole niihin pääsyä (Lee et al. 2006, s. 84).

Datan laadun raporttien julkaiseminen ja jakaminen asiaankuuluville sidosryhmille tarkoittaa, että datan laadun ulottuvuudet ymmärretään laajalti (mikä auttaa hälventämään

mahdollisia virheellisiä käsityksiä laadusta). Raportit voivat toimia myös eräänlaisena kannustimena laadunparannuksiin, sillä datan laadun näyttäminen organisaation eri alueilla tuo kilpailukyvyn elementin datan laadunhallintaan. Ajan mittaan raporteilta voi myös seurata laadun muutosta, mikä auttaa varmistamaan, että implementoiduilla laadunparannusratkaisuilla on vaikutusta ja lähestymistapaan tehdään muutoksia, mikäli laatu alkaa heikkenemään. (Schwarzenbach & King 2020, s. 74)

3.3 Roolit ja vastuut

Dataan ja sen (laadun)hallintaan liittyviä toimia on tyypillisesti pidetty puhtaasti IT-organisaation tehtävänä, vaikka datan laadunhallinta on organisaatiota poikkileikkaavaa toimintaa. Data tuotetaan liiketoimintaprosesseissa, joihin IT:llä ei ole näkyvyyttä tai kosketuspintaa. (Sebastian-Coleman 2013, s. 18). Vaikka IT ei omista liiketoimintaprosesseja, heidän vastuullansa on valvoa datan laatua esimerkiksi datan käsittelyn eri vaiheissa, kuten datan muokkaamisen ja siirtymisen aikana ja omalta osaltaan pitää huolta tarvittavista laatutoimenpiteistä, sekä pystyttää datan laadun mittaamiseen ja analysoimiseksi tarvittava teknologia (Sebastian-Coleman 2013, s. 22–23). Loppukäyttäjät voivat toimia viimekädessä laadunvarmentajana tutkiessaan datasta tuotettuja raportteja asiantuntijuutensa vuoksi (Treder 2020, s. 150). Kuitenkaan laatuvirheitä ei ensisijaisesti pitäisi päätyä raporteille käyttäjien tarkistettavaksi, vaan laatua tulee kehittää proaktiivisesti, iteratiivisesti ja ongelmat havaita mahdollisimman aikaisessa vaiheessa datan elinkaarta.

Datan elinkaaren hallintaan liittyy useita rooleja, jotka karkeasti voidaan jakaa kolmeen: datan kerääjä/tuottaja (data collector/producer), datan käsittelijä (data custodian) sekä datan käyttäjä (data consumer) (Lee & Strong 2003, s. 17; Mahanti 2019, s. 449; Sebastian-Coleman 2013, s. 19–20). Kerääjät tai tuottajat tyypillisesti ovat lähdejärjestelmäpäässä tietoja syöttäviä tahoja, joita voidaan myös kutsua datan luojiksi (data creator) (Treder 2020, s. 130). Datan tuottajat voivat olla myös tietojärjestelmiä, organisaatioita tai muita tahoja, jotka luovat dataa (Mahanti 2019, s. 449). Datan käsittelijä voi olla esimerkiksi datainsinööri- tai analyytikko, joka tuntee datan käsittelyyn tarkoitetut työkalut ja järjestelmät, kun taas datan käyttäjä usein on loppukäyttäjä, eli esimerkiksi liiketoiminnan asiantuntija, joka tuntee datan sisällön (Loshin 2011, s. 27).

Muita datan hallinnan rooleja ovat esimerkiksi datan omistajat (data owner) sekä datavastaavat (data steward) (Sebastian-Coleman 2013, s. 20–21) ja monet muutkin roolit, jotka ovat mukana datan syöttöprosessissa tai esimerkiksi määrittämässä sitä, minkälainen datan rakenne tietojärjestelmässä tulee olla tai miten se tulee syöttää järjestelmään, jotta se on oikein (Loshin 2011, s. 28–29). Datavastaavan tehtävänä on varmistaa, että datan laatu täyttää vaatimuksensa tietyn data-alueen osalta (Mahanti 2019, s. 449). Datan omistaja puolestaan vastaa tietystä data-alueesta ja hänellä on päätösvalta tätä dataa koskevissa päätöksissä (Mahanti 2019, s. 452). Kaikilla on oma vastuunsa siitä, minkälaista data on sen elinkaaren eri vaiheissa, joka vahvistaa kokonaisvaltaisen laadunhallinnan sanomaa siitä, että laatu on kaikkien vastuulla ja datan laatua tulee hallita koko datan toimitusketjun ajan (Treder 2020, s. 151, s. 171).

Datan laadunhallinnassa on erilaisia prosesseja ja aktiviteetteja, sekä kuten mainittu, myös erilaisia rooleja, jotka ovat kukin vastuussa omista tehtävistään. Tätä tehtävänjakoa voidaan visualisoida ja kartoittaa RACI-mallin avulla, johon kerätään olennaisia datan laadunhallinnan tehtäviä ja niistä vastuussa olevia rooleja (Loshin 2011, s. 85). Taulukossa 6 on esitetty esimerkki RACI-mallista datan laadunhallintaan. Taulukko 6: RACI-vastuunjakotaulukko datan laadunhallintaan (mukaillen Loshin 2011, s. 87-88). RACI-malli on kaksiulotteinen matriisi, jossa tehtävät on listattu riveittäin ja sarakkeet edustavat rooleja. Jokaiseen soluun merkitään kirjain, joka kuvaa roolin vastuuta kyseisestä tehtävästä. Kirjaimet ovat R (responsible) = vastuullinen tehtävän suorittamisesta, A (accountable) = päätöksentekijä/vastaava, joka pitää huolen, että tehtävä tulee tehdyksi, C (consulted) = tältä roolilta kysytään mielipidettä, neuvoa, apuja tehtävien tekemiseen, tai I (informed) = tätä roolia tiedotetaan ja pidetään ajan tasalla tehtävän edistyessä. Kaikkia sarakkeiden rooleja ei luonnollisesti välttämättä löydy organisaatiosta, jolloin tehtäviä on hyvä sovittaa olemassa oleviin rooleihin. (Loshin 2011, s. 85–86)

4 Toiminnan kuvaaminen

Tässä kappaleessa käsitellään tutkimusm Helenin toimintaympäristöä ja organisaatiota, datan omistajuusrakennetta, salkkuennusteprosessia, sekä sen käyttämää dataa, nykyisiä laadunhallintatoimenpiteitä- ja työkaluja. Toiminnan kuvausta varten käytiin työpajoja sekä muutamia haastatteluja, jotka ovat esitetty taulukossa 7 alla.

Taulukko 7: Vaihe 1 - Nykytilan ja liiketoimintavaatimusten kartoitus

Nro	Osallistuja(t)	Menetelmä	Aihe	Tarkoitus	Pvm
1	Data ja tekoäly-yksikkö, toimittajan konsultit	Teams-työpaja	Datan laadun viitekehyksen läpikäynti ja vaatimusten kartoitus	Toimittajan viitekehyksen läpikäynti	24.2.2022
2	Data ja tekoäly-yksikkö, toimittajan konsultit	Teams-työpaja	Datan laadun viitekehyksen läpikäynti ja vaatimusten kartoitus	Tekninen läpikäynti - arkkitehtuuri, teknologia ja toimintatavat	14.3.2022
3	Kehityspäällikkö	Teams-haastattelu	Asiakasdata ja sopimustietojen kirjaus	Nykyisten laadunhallintamenetelmien kartoitus	15.3.2022
4	Kehityspäällikkö	Teams-haastattelu	Asiakaspalvelun ja sopimustietojen virhelistat	Nykyisten laadunhallintamenetelmien kartoitus	15.3.2022
5	Data ja tekoäly-yksikkö, sähkökaupan asiantuntija, toimittajan konsultit	Teams-työpaja	Datan laadun viitekehyksen läpikäynti ja vaatimusten kartoitus	Liiketoimintavaatimusten kartoitus - prosessi, toimintatavat ja roolit sekä vastuut	28.3.2022

Haastattelujen ja työpajojen lisäksi toiminnan kuvaamisessa on hyödynnetty Helenin omia materiaaleja syventämään ymmärrystä prosesseista, tietovirroista ja nykyisistä laadunhallintamenetelmistä. Materiaalit on kuvattu taulukkoon 8.

Taulukko 8: Viitekehyksen jalkautuksen tukena ja ymmärryksen syventämiseen hyödynnetyt Helenin sisäiset materiaalit

Nro	Nimi	Kuvaus
1	Salkkuennuste - järjestelmäkuvaus	PowerPoint
2	Salkkuennusteet yhteenveto	PowerPoint
3	Salkkuennuste tarkastuslista	PowerPoint
4	Pienasiakasmyynti salkkuennuste vastuut	PowerPoint
5	salkkuennuste_kuvaus	PowerPoint
6	Salkkuennuste_DiRa_3_2022	PowerPoint
7	Sähkö - Pienasiakkaat salkkuennuste-raportti	PowerBI-raportti
8	Datan omistajuusrakenne Helenissä	PowerPoint
9	Toimittajan dokumentaatio teknisestä DQ-moduulista	ZIP-tiedosto, useita tiedostoja (.py, .html, .docx, ym.)

4.1 Toimintatutkimuksen olemus

Toimintatutkimus keskittyy sekä kehitystyöhön että tutkimukseen, eli se ei varsinaisesti ole tutkimusmenetelmä vaan lähestymistapa. Tarkoituksena on saada aikaan ongelmien ratkaisuun ja muutoksiin johtavaa toimintaa, jossa yhdistyy sekä perinteisiä tutkimuksia ohjaava teoreettinen intressi (miten asiat ovat), että käytännöllinen intressi (miten asiat voisi tehdä paremmin). Näiden intressien lisäksi pohditaan, miten tahtotilaan päästään ja mitä tietoa sen saavuttamiseksi tarvitaan. Se on myös luonteeltaan sosiaalista, eli keskeistä on käytännöissä mukana olevien ihmisten osallistuminen ja osallistaminen. Toisin sanoen yksi

toimintatutkimuksen pääpiirteistä liittyy tutkijan ja organisaation jäsenen väliseen yhteistyöhön organisatoristen ongelmien ratkaisemiseksi. (Valli 2019)

Seuraavat toimintatutkimuksen piirteet on otettava huomioon harkittaessa sen soveltuvuutta mihin tahansa tutkimukseen:

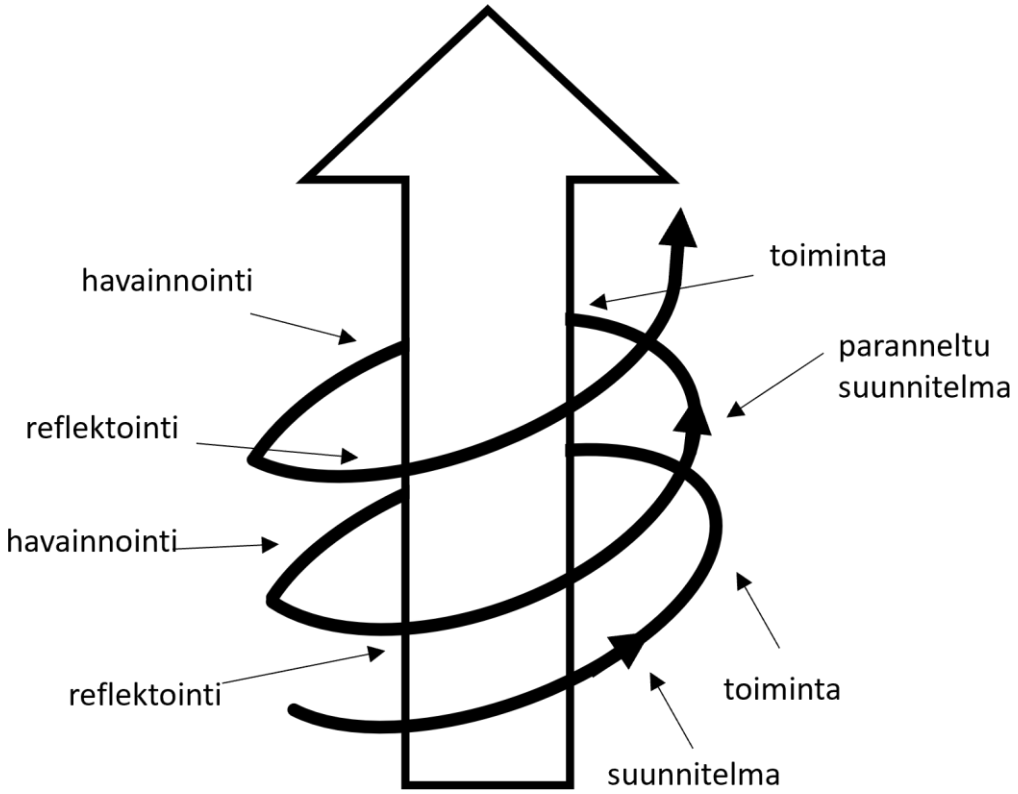
- Sitä sovelletaan tiettyjen käytäntöjen parantamiseksi (Suojanen 2014).
- Toimintatutkimus perustuu kerättyyn tietoon perustuvaan toimintaan, arviointiin ja kriittiseen käytäntöjen analysointiin, jotta asiaankuuluviin käytäntöihin saadaan parannuksia (Valli 2019).
- Toimintatutkimusta helpottaa useiden henkilöiden osallistuminen ja yhteistyö, joilla on yhteinen tarkoitus.
- Toimintatutkimus keskittyy tiettyihin tilanteisiin ja niiden kontekstiin.

Toimintatutkimuksen nähdään syntyneen Kurt Lewinin (1890–1947) työn tuloksena. Lewin (1944) esitti näkemyksensä osallistumisesta ulkoisena toiminnan tutkijana organisaatiomuutokseen, ja kaksi hänen lähimmistä työtovereistaan, Alfred Marrow ja John French, kuvailivat, kuinka he osallistuivat toimintatutkijoina mahdollistamaan muutoksen toteutumisen (Willis & Edwards 2014, s. 30). Toimintatutkimus olettaa, että sosiaalinen maailma muuttuu jatkuvasti, ja sekä tutkija että tutkimus ovat osa tätä muutosta. Tätä silmällä pitäen, tutkijan rooli on erilainen kuin muissa tutkimusmenetelmissä, sillä toimintatutkimuksessa hänen on oltava tutkimusryhmän aktiivinen jäsen koko tutkimuksen ajan. Myöskään lopputuloksia ei voida jättää täysin ulkopuolisten arvioitavaksi, vaan tutkimuksen osalliset kontrolloivat kaikkia tutkimusprosessin vaiheita (Suojanen 2014). Aktiivinen osallistuminen kehitystyöhön edellyttää, että tutkijalla on ennestään ymmärrystä ja kokemusta käsiteltävästä aiheesta sekä kohdeorganisaatiosta (Juuti & Puusa 2020).

Yleisesti toimintatutkimukset voidaan jakaa kolmeen kategoriaan: positivistinen, tulkitseva ja kriittinen. Positivistinen lähestymistapa toimintatutkimukseen, joka tunnetaan myös nimellä "klassinen toimintatutkimus", näkee tutkimuksen sosiaalisena kokeiluna. (Willis & Edwards 2014, s. 24–25) Tulkitseva toimintatutkimus näkee liiketoimintatodellisuuden sosiaalisesti rakentuneena ja keskittyy tutkimuksessa eritoten paikallisten ja organisatoristen tekijöiden määrittelyyn (Willis & Edwards 2014, s. 31). Kriittinen toimintatutkimus on

erityinen toimintatutkimuksen muoto, joka suhtautuu kriittisesti liiketoimintaprosesseihin ja tähtää parannuksiin (Willis & Edwards 2014, s. 36).

Toimintatutkimusprosessi sisältää erilaisten vaiheiden muodostamia syklejä. Useimmat toimintatutkimusmallit noudattavat tätä kaavaa, eli sisältävät neljästä kuuteen vaihetta, joita seurataan iteratiivisesti spiraalin muodossa (Willis & Edwards 2014, s. 59). Päävaiheet ovat tyypillisesti suunnitelma, toiminta, havainnointi ja reflektointi (Zuber-Skerritt & Wood 2019). Tätä ideaa on havainnollistettu myös Kurt Lewinin toimintatutkimuksen perusmallissa kuvassa 8 alla.



Kuva 8: Toimintatutkimuksen perusmalli (mukaillen Willis & Edwards 2014, s. 13)

Prosessi lähtee liikkeelle ongelmien tunnistamisesta toiminnan suunnitteluun, toimintaan ja havainnointiin, jonka jälkeen tuloksia reflektoidaan. Reflektointivaiheessa keskustelut kollegoiden ja yhteistyökumppaneiden kanssa luovat yhteisymmärrykseen siitä, mitä ongelmat ovat ja mitä toimintatutkimuksen tavoitteena tulisi olla (Valli 2019). Tässä vaiheessa ongelmien määritys ja tavoitteiden asetanta tyypillisesti on vielä alustavaa ja voi muuttua monta kertaa tutkimuksen aikana. Kun tavoitteesta on muodostettu alustava käsitys, seuraavaksi

tutkitaan tapoja ratkaista se ongelma, joka on syy tutkimuksen tekemiseen. Tämä vaihe voi sisältää monenlaisia aineistonkeruumuotoja - kirjallisuuskatsauksista ja konferensseihin osallistumisesta kollegoiden ja asiantuntijoiden kuulemiseen haastattelujen tai työpajojen muodossa, sekä vertailujohtamalla (benchmarking) muita tapauksia, joissa on onnistuneesti ratkaistu tai käsitelty vastaavanlaisia ongelmia (Tietoarkisto, n.d.). Lopulta tutkimusryhmä keskustelee vaihtoehdoista ja laatii suunnitelman ongelman ratkaisemiseksi. Tämän jälkeen suunnitelmaa lähdetään toteuttamaan sellaisilla resursseilla, tuella ja ryhmäkokoonpanolla, joita suunnitelma vaatii onnistuakseen. Toteutuksen jälkeen muutosta tarkkaillaan ja sitä arvioidaan huolellisesti. Muutoksen tehokkuutta ja onnistumista arvioidaan erilaisilla laadullisilla ja määrällisillä keinoilla. Usein muutoksen arviointi on epävirallista, laadullista ja perustuu osallistujien väliseen pohdiskeluun ja dialogiin. Kriittisessä toimintatutkimuksessa voidaan hyödyntää myös systemaattista tulosten keräämistä, mutta positiivisen toimintatutkimuksen mallissa tätä ei yleensä tehdä. Syklien tuloksien avulla suunnitelmaa tyypillisesti muotoillaan uudelleen, sillä toimintatutkimus harvemmin päättyy yhteen sykliin. Tämä johtuu siitä, että muutosta harvoin saadaan jalkautettua onnistuneesti ensimmäisellä kerralla. Syklejä toistetaan, kunnes tutkimusryhmä on tyytyväinen saavuttamiinsa tuloksiin (Tietoarkisto, n.d.). (Willis & Edwards, 2014, s. 59–60)

4.2 Helen organisaatio ja toimintaympäristö

Helen on yksi Suomen suurimmista energiakonserneista, jonka pääliiketoimintaan kuuluu sähkön, kaukolämmön ja -jäähdytyksen tuotanto sekä energian jakelu ja myynti. Konsernin tytäryhtiö Helen Sähköverkko Oy vastaa sähkön siirto- ja jakelupalveluista lähes koko Helsingin alueella. Helenin asiakaskunta koostuu 550 000 yksityis- sekä yritysasiakkaasta ympäri Suomea. Yrityksen strategiset tavoitteet sekä missio olla hiilineutraali energiayhtiö vuonna 2030 ovat kiihdyttäneet tarvetta kehittää älykkäämpiä ratkaisuja paremmin vastaamaan asiakkaiden tarpeisiin. Kehittääkseen älykkäämmän hiilineutraalin energiajärjestelmän, jolla voidaan tuottaa, käyttää ja säästää energiaa ympäristöä varten, tarvitaan skaalautuvaa ja tietoturvallista IT-infrastruktuuria, yhteisiä ja ketteriä toimintatapoja sekä liiketoiminnan ja asiakkaiden tarpeisiin vastaavia sovelluskokonaisuuksia. (Helen Oy 2020a)

Digitaaliset Ratkaisut (DiRa)- organisaatioyksikkö vastaa Helenin kaikkien digitaalisten ratkaisujen käytettävyydestä, saatavuudesta ja kehityksestä. Tähän kuuluu myös vastuu

Helenin asiakkaiden digitaalisten kanavien ja palveluiden kehittämisestä ja ylläpitämisestä, siten että kanavat muodostavat eheän kokonaisuuden ja tukevat erinomaisessa asiakaskokemuksessa. Digitaalisten ratkaisujen tehtävänä on huolehtia datan hyödynnettävyydestä ja sen jalostamisesta johtamisen, liiketoiminnan ja asiakkaiden hyödyksi. DiRa jakautuu viiteen eri yksikköön, joista Data- ja tekoäly-yksikkö on rakentamassa data-alustaa, jonka päällä suuri osa digitaalisista palveluista toimivat. Data- ja tekoäly-yksiön tehtäviin kuuluu myös kehittää yhteisiä toimintatapoja ja hallintamenetelmiä datan käytön osalta organisaatiolaajuisesti. (Helen Oy 2022a)

Helenissä toimii analytiikan yhteistyöverkosto, joka koostuu kolmesta eri tasosta. Strategisesta päätöksenteosta vastaa talousjohtaja, joka edustaa tehtävässään Helenin johtoa. Kehitystarpeiden katselmoinnista vastaa analytiikan ohjausryhmä, mikä koostuu eri liiketoimintojen ja kumppaneiden edustajista. Operatiivisen toiminnan vastaavana on liiketoiminta sekä Data ja tekoäly-yksikkö yhdessä ketterinä virtuaalitiimeinä, joita kutsutaan heimoiksi. Kullakin liiketoiminta-alueella on oma heimonsa, jonka toimintaa koordinoi vastuuanalyytikko yhdessä liiketoiminnan edustajan sekä heimon jäsenten kanssa. Data ja tekoäly-yksikön ”Datan hallinta ja kyvykkyydet”-tiimi huolehtii yhdessä liiketoimintojen, liiketoimintakumppanien ja datan omistajien kanssa siitä, että tarvittavat tiedon laadun mittarit on määritelty ja niiden seuranta on järjestetty. Kyseinen tiimi myös viestii tiedon laadun merkityksestä muulle organisaatiolle ja pyrkii vaikuttamaan virheiden juurisyihin, muun muassa kehittämällä ymmärrystä datan laadusta organisaatiolaajuisesti erilaisten verkkokurssien ja sisällöntuottamisen avulla. (Helen Oy 2020b)

4.3 Datan laadun nykytila

Data ja tekoäly-yksikkö aloitti vuoden 2021 kesällä kartoittamaan datan laadun nykytilaa, jonka pohjalta kesän päätteeksi laadittiin toimintasuunnitelma ja suoritettiin kokeiluja datan laadun mittaamiseksi. Nykytila-analyysissa nousi esille erilaisia puutteita ja kehitysehdotuksia, joista osa ovat korjautuneet uusien tietojärjestelmäimplementaatioiden ja vanhojen järjestelmien alasajojen myötä. Esille nousi myös positiivisia havaintoja siitä, miten hyvällä mallilla tietyissä liiketoimintaprosesseissa (kuten myynnin ja asiakaspalvelun osalta) laadun seuranta myös on. Datan laadunhallintamenetelmiä ei kuitenkaan vielä laajemmin otettu syksyn 2021 aikana käyttöön uudella data-alustalla, vaan datan laadunhallintaviitekehyksen

kehittämiseen ja jalkauttamiseen päätettiin ottaa ulkopuolinen toimittaja sekä diplomityöntekijä avuksi.

4.4 Sähkön suojaus ja salkkuennusteet

Helenin asiakkaille on tarjolla monenlaisia ja erihintaisia sopimuksia, jotta kuluttajilla on mahdollisuus valita omaan elämäntilanteeseensa sopiva vaihtoehto. Pörssisähkösopimuksessa hintariski on isompi asiakkaalla, kun taas kiinteähintaisessa Helen ottaa riskin. Kiinteähintainen määräaikainen sähkösopimus myydään asiakkaalle kiinteään hintaan tyypillisesti kahdeksi vuodeksi. Helenin tuottama ja osuuslaitoksista hankkima sähkö myydään sähköpörssiin, kun taas sähköasiakkaille myytävä sähkö ostetaan sähköpörssistä, jotta voidaan tarjota erityyppisiä sopimuksia myös omien voimaloiden tuotantokapasiteetin ulkopuolelta.

Suojauskaupassa sähkön hinta suojataan etukäteen tietylle ajanjaksolle, jotta ennusteiden avulla voidaan esimerkiksi myydä osa sähköstä vuoden päähän. Näin sähkön perushinta saadaan varmistettua pitkälle tulevaisuuteen ja asiakkaille voidaan tarjota kiinteähintaisia sopimuksia. Salkkuennusteen avulla pyritään suojaamaan kotien ja pienyritysten sähkönostot. Tämä kattaa määräaikaiset, toistaiseksi voimassa olevat sekä markkinahintasopimukset.

Suojaustaso tyypillisesti toteutetaan kulutusennusteiden mukaan, eli mikäli todellinen kulutus on suurempi kuin ennustettu kulutus, ylimääräinen osuus joudutaan hankkimaan spotmarkkinoilta markkinahintaan, joka voi vaihdella suuntaan tai toiseen sähkön myyntihinnasta. Sähköpörssissä sähkön hinta on erittäin volatiili, sillä hintaan vaikuttavat sää- ja luonnonolosuhteet, kulutusnäkymät, muutokset polttoaineiden ja päästöoikeuksien hinnoissa ja tuotantotilanne. Myös lainsäädäntö määrää sähkömarkkinoiden hinnoittelua omalta osaltaan, eli sähkömarkkinalaki rajoittaa muun muassa hinnannousua vuositasolla, jonka vuoksi muutoksia on seurattava ja tehdä päätöksiä siitä, miten sähköä ja polttoaineita kannattaisi ostaa ja myydä.

Hinta on myös kuluttajalle usein sähkösopimuksessa merkittävä tekijä. Etenkin viime aikoina mediassa on käyty paljon keskustelua siitä, onko pörssisähkösopimus vai määräaikainen sähkösopimus kuluttajalle edullisempi vaihtoehto aiemman sopimuksen päättyessä, sillä hinta on voinut nousta edellisestä sopimuksesta huomattavasti edellä mainituista syistä.

(Helen Oy 2022b)

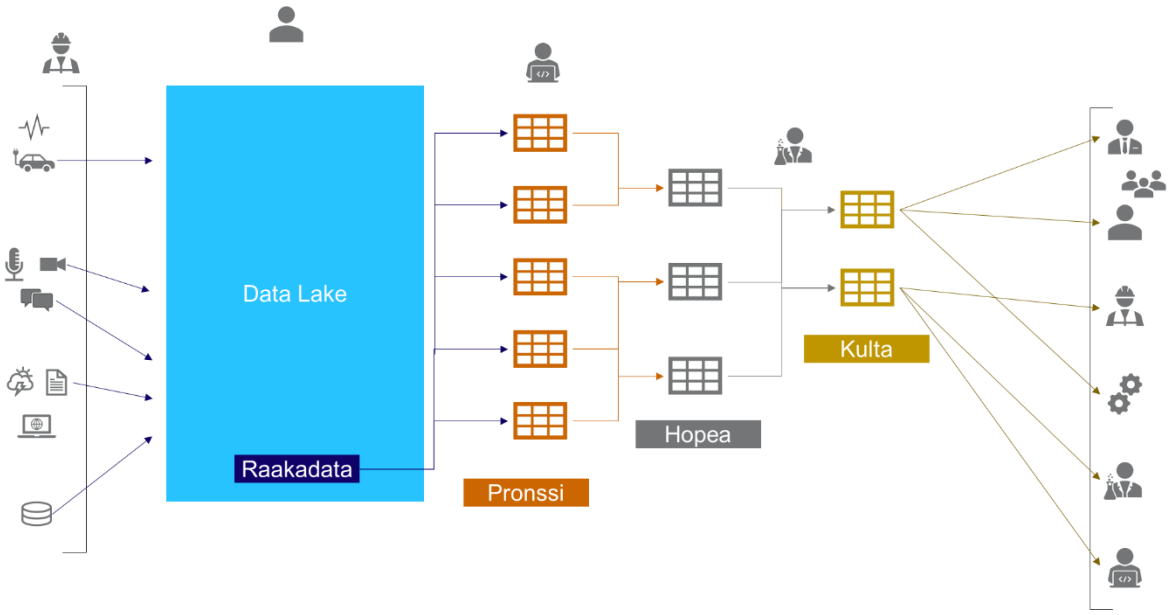
4.5 Prosessin kuvaus

Salkkuennusteprosessissa käytettävä data voidaan karkeasti jaotella muutamaan tietoalueeseen, jotka ovat sopimustiedot, ennakkotilaukset, odotetut uusintatilaukset sekä lämmitystarveluvut. Prosessi toimii niin, että aina tiettynä viikonpäivänä data ladataan sisään salkkuennusteraportille, jonka raportista vastaava asiantuntija validoi manuaalisesti tarkistusraporttien ja tarkistuslistan avulla. Tärkeimmät tarkistettavat asiat ovat ennuste vs. lähdedata, ennustevirhe ja ennusteen performanssi sekä vertailumuutos edelliseen viikkoon. Tarkastuslistalla on yhteensä kuitenkin seitsemän tarkastusvaihetta. (Taulukko 9, numero 6)

Tämän jälkeen sähkökaupan asiantuntija Myynti- ja asiakaspalvelu (MAP) -liiketoiminnosta tarkistaa ennusteen ja vertailee muutosta edelliseen viikkoon, jonka jälkeen ennuste viedään energiatietojen hallintajärjestelmään. Prosessin viimeisessä vaiheessa energianhallinta ja tukkukauppa (EHT) -liiketoiminnan asiantuntija tekee tarvittavat hintasuojaukset (ostoa tai myyntiä) saamansa datan sekä sisäisten ohjeistuksien mukaisesti. (Taulukko 9, numero 6)

4.6 Data-alustan arkkitehtuuri ja teknologiat

Data-alustan arkkitehtuuri on rakennettu niin kutsuttuun mitaliarkkitehtuurimuotoon, eli se sisältää raaka, pronssi, hopea sekä kultakerrokset. Tavoitteena on asteittain parantaa datan rakennetta ja laatua, kun se virtaa arkkitehtuurin jokaisen kerroksen läpi. Kaikki data on data-alustalla deltatauluissa, joilla on luontaisesti joitain parempia ominaisuuksia kuin perinteisillä tauluilla, jotka tallentavat rivi- ja sarakemuodossa. Helenin data-alustan arkkitehtuuri on esitetty kuvassa 9 alla.



Kuva 9: Data-alustan arkkitehtuuri

Lähdejärjestelmistä data tulee pitkälti raakakerrokseen sellaisenaan, eli datan rakenne pyritään pitämään yksi yhteen datalähteen kanssa. Kaikki raakadata kerätään lähtökohtaisesti tänne tiedostomuodosta riippumatta, eli data saattaa olla esimerkiksi csv-, excel, json tai xml-tiedostoina. Pronssikerrokselle vietäessä raakadata muunnetaan eri tiedostoformaateista yhteiseen muotoon, eli niin kutsuttuun deltamuotoon. Tässäkin kerroksessa rakenne vastaa vielä lähdejärjestelmää, mutta dataan voidaan lisätä teknistä lisätietoa, kuten aikaleima- ja datalähdetietoa ja se pyritään kuvaamaan mahdollisimman tarkalla tasolla, esimerkiksi mittaridata minuuttitasolla. Hopeatason datalle tehdään jo enemmän käsittelyä, sillä tässä kerroksessa dataa siivotaan, sarakenimiä yhtenäistetään, dataan yhdistetään muita tietolähteitä ja sitä voidaan jalostaa liiketoimintalogiikalla. Rakenne pyritään edelleen pitämään samalla tasolla kuin lähdejärjestelmässä, eikä dataa ei myöskään aggregoida vielä, vaan se pidetään mahdollisimman tarkalla tasolla. Kultakerroksessa dataa mallinnetaan enemmän ja se pitää sisällään esimerkiksi raportointia varten valmiimmaksi jalostettua tietoa. Rakenteita muunnetaan enemmän liiketoimintalähtöiseksi ja lähdejärjestelmäriippuvaisuus pienenee. Tietoja on esimerkiksi pilkottu useampiin eri käsitteisiin liittyviin tauluihin. Dataa on mahdollisesti summattu ylemmälle tasolle edellisten kerrosten tarkemmasta tasosta, esimerkiksi mittaridataa voidaan summata tuntitasolle, kun taas hopeakerroksella mittaridata voi olla minuuttitasolla.

Data-alustan taustalla olevat teknologiat ovat pitkälti Microsoftin (Azuren) palveluita. Dataputkia orkestroidaan Synapse Analytics-työkalulla. Työkalussa voidaan hallinnoida datan

siirtämiseen ja muokkaamiseen tarvittava transformointi, dataputkien ajastukset, ajojen seurannat ynnä muut tyypilliset ETL (extract-transform-load) -toiminnallisuudet. Varsinainen laskentalogiikka tapahtuu Databricksin työkirjoilla, jotka ovat siis python, SQL, R, tai scala-ohjelmointikielillä kirjoitettuja työkirjoja. Nämä sisältävät erilaisia toimintoja, joilla dataa muokataan haluttuun muotoonsa yllä olevan arkkitehtuurin eri kerroksien vaatimalla tavalla.

Eräänlaisena yhteistyö- tai projektinhallintatyökaluna Data ja tekoäly-yksikkö sekä liiketoiminnot käyttävät Azure Devopsia. Sen avulla voidaan kehittää ja parantaa tuotteita nopeammin kuin perinteisillä ohjelmistokehitysmenetelmillä. Työkaluun voi esimerkiksi kirjata käyttäjätarinoita, tehtäviä, bugeja, joille voidaan asettaa tavoiteaikataulu ja niiden edistymistä voidaan seurata.

Dataputkien ajoja sekä muuta lokitietoa voidaan seurata Azure Log Analyticsin avulla. Se on työkalu, jolla voi muokata ja suorittaa lokikyselyitä Azure Monitor -lokien keräämistä tiedoista ja analysoida niiden tuloksia interaktiivisesti. Kyselyjen perusteella voidaan tunnistaa trendejä, analysoida malleja ja laatia johtopäätöksiä datasta. Azure Monitor Alerts puolestaan on eräänlainen hälytysten valvontatyökalu. Hälytyksien avulla voidaan havaita ja ratkaista ongelmia, ennen kuin virhe päätyy loppukäyttäjille saakka.

4.7 Nykyiset laadunhallintamenetelmät

Työpajoissa käytiin läpi kysymysten ja keskustelun avulla nykyisiä laadunhallintamenetelmiä salkkuennusteprosessissa sekä tarpeita uusia laadunhallintamenetelmiä ajatellen. Keskusteluissa kävi ilmi, ettei Helenillä ei ollut kyseiseen prosessiin olemassa varsinaisia datan laadun tarkistuksia projektin alkaessa. Ennustevirhettä ja tiettyjä yksittäisiä sisältötarkistuksia kuitenkin seurataan, kuten käyttöpaikkojen lukumäärää, lämpötilatietoja, tuotetietoja ja tilaus- ja tarjoustietoja (taulukko 3, numero 4). Näille on määritetty ohjeet (taulukko 3, numero 3), kuinka ne validoidaan ja niitä seurataan manuaalisesti analyytikon sekä sähkökaupan asiantuntijan puolesta kerran viikossa pääsääntöisesti PowerBI-raportilta. Teknisessä läpikäynnissä nousi esille, että tyypillisesti virheiden esiintyessä virheen havainnut henkilö lähettää sähköpostia osallisille, jota kautta ongelma sitten ratkotaan. Toisinaan virheistä luodaan myös tiketti Azure Devops-työkaluun, mikäli kyseessä on esimerkiksi datainsinöörin havaitsema bugi koodissa, joka vaatii dataputkien korjausta.

Salkkuennusteprosessin vastuuanalyytikko on arvioinut, että prosessiin käytetyn datan huono laatu voi aiheuttaa suuriakin taloudellisia haittoja, sillä virheellisillä päätöksillä saatetaan sähköä suojata ja ostaa väärin, mikä näkyy tällaisessa volyymikaupassa isoina tulonmenetyksinä.

Asiakas- ja sopimustietohaastattelusta selvisi, että lähdejärjestelmäpuolella datan laatua seurataan automaattisesti erilaisten SQL-hakujen avulla ja erityisesti kirjauskäytäntöihin on panostettu. SQL-hakuja on laadittu perustuen liiketoiminnan tarpeisiin ja näistä lähtee myös korjauksista vastaavalle sopimustenhallintayksikölle ajastetusti lista korjattavista virheistä. Näin varmistetaan, että korjaukset tehdään lähdejärjestelmään oikein, sen sijaan, että dataa imputoitaisiin tai manipuloitaisiin data-alustalla. Kirjauskäytännöistä keskustellaan aktiivisesti MAP-heimon Teams-ryhmässä, jonne myös ohjeistuksia kerätään. Erään laajan järjestelmähankkeen myötä datan laatuun on myös keskitytty huomattavasti enemmän, sillä muun muassa tiettyjen kenttien tarkkuusvaatimukset ovat olleet edellytyksenä projektin läpiviennille.

Materiaaleista ja läpikäynneistä kävi myös ilmi, että datan profilointia on tehty pandas-profiling python-kirjaston avulla kaikkiin data-alustan kultatason tauluihin, jotka on sittemmin linkitetty datakatalogiin kyseisten datalähteiden tietoihin. Raaka-, pronssi-, tai hopeatason taulujen profiileja ole luotu, joten datan profiili näiden kerrosten taulujen osalta ei ole tiedossa. Varsinaisia datan laatutarkistuksia, kuten rivimäärä- tai tuoreustarkistuksia on myös joidenkin tietolähteiden osalta toteutettu deequ python-kirjastolla, joista lähtee automaattinen Teams-hälytys testien palautuessa virheellisinä. Nämä ovat todettu erityisesti datainsinöörien kannalta hyödyllisiksi, sillä ne nopeuttavat ongelmanratkaisuprosessia sekä lisäävät läpinäkyvyyttä. Salkkuennusteeseen käytettävään dataan näitä ei ole kuitenkaan vielä monistettu.

Työpajojen perusteella toivottiin myös tarkempaa tietoa dataputkien terveydestä, eli niiden läpimenosta ja kestosta. Lisäksi haluttiin kyetä seuraamaan datan sisällöllistä laatua, kuten täydellisyyttä ja tiettyjen kenttien validiteettia. Näiden tietojen avulla päästäisiin helposti kiinni siihen, missä virhe on tapahtunut. Tarkistuksien ohelle olennaisina toiminnallisuuksina nähtiin hälytykset virheistä sekä dataongelmien tiketöintityökalu, mutta mahdollisimman pitkälle automatisoituna ja niin, ettei se muuta nykyisiä toimintatapoja paljon, jottei aikaa kulu manuaaliseen selvitys- ja seurantatyöhön. Esille nousi myös keskustelua

poikkeamien tunnistuksesta koneoppimisen avulla, jota pohdittiin toteutettavan Azuren Metrics Advisor palvelulla.

Toteutukseen toivotut toiminnallisuudet ovat kerätty taulukkoon 9 alla, sekä prioriteetti, joiden mukaan projektissa edetään. Mikäli projektissa ilmeni esteitä aikataulun tai teknisten rajoitteiden osalta, otettiin huomioon kyseisen toiminnallisuuden prioriteetti ja sen tarpeellisuutta uudelleenarvioitiin.

Taulukko 9: Datan laadunhallinnan toteutukseen toivotut toiminnallisuudet

Nro	Toiminnallisuus	Kuvaus	Prioriteetti
1	Dataputkien terveys	Tekniset mittarit, eli kuinka kauan ajot kestävät, kasaantuvatko ajot tietyille ajanjaksoille, ajot statuksien (succeeded, failed, jne.) mukaan.	1
2	Datan sisällölliset mittarit ja niiden tallennus	Datalle määritetään sisällöllisiä mittareita ja mittauksista syntyvät tulokset pystytään tallentamaan tietokantaan.	1
3	Datan laadun raportointi	Mittareiden tuottama data visualisoidaan eri käyttötarpeisiin (liiketoiminta ja tekninen näkökulma).	1
4	Hälytykset ja tiketöinti	Mittareille määritetään raja-arvot, joiden mukaan hälytyksiä ja tikettejä syntyy. Hälytykset Teams-viestinä, tiketöinti Azure Devopsiin.	2
5	Datan elinkaari	Datan elinkaaren visualisointi (liikkuminen alkulähteeltä loppuraportille).	3
6	Profilointi	Muiden taulujen profilointi eri mitaliarkkitehtuurin tasoilla (raaka, pronssi, hopea) olemassa olevalla työkalulla.	3
7	Poikkeamien (anomaly) tunnistaminen	Poikkeamien tunnistaminen koneoppimisen avulla (Metrics Advisor).	3
8	Datan laadunhallintaan kuluvien resurssien seuranta	Devops tiketteihin kuluva aika, paljonko auki / kiinni tikettejä, ym.	3

5 Uusien laadunhallintamenetelmien käyttöönotto

Uusien datan laadunhallintamenetelmien avulla nykyistä prosessia voidaan automatisoida, virtaviivaistaa sekä selkeyttää prosessissa toimivien tahojen vastuita ja tehtäviä, mutta myös parantaa datan laatua ja sitä kautta ennustetarkkuutta. Mitä aiemmin virheisiin päästään kiinni ja laatuongelmat henkilöityä oikeille tahoille, sitä pienempi taloudellinen riski tai vaikutus virheillä on. Menetelmät ovat rakennettu niin, että niitä voidaan tulevaisuudessa monistaa toisille alueille tekemällä pieniä muutoksia, sillä ne pohjautuvat teknologiaan, jota voi skaalata ja mukauttaa tarpeiden vaatiessa. Menetelmien laajentaminen muille alueille tosin vaatii aina aluksi liiketoimintatarpeiden määrittämisen, kuten kaikissa datan laadun hankkeissa.

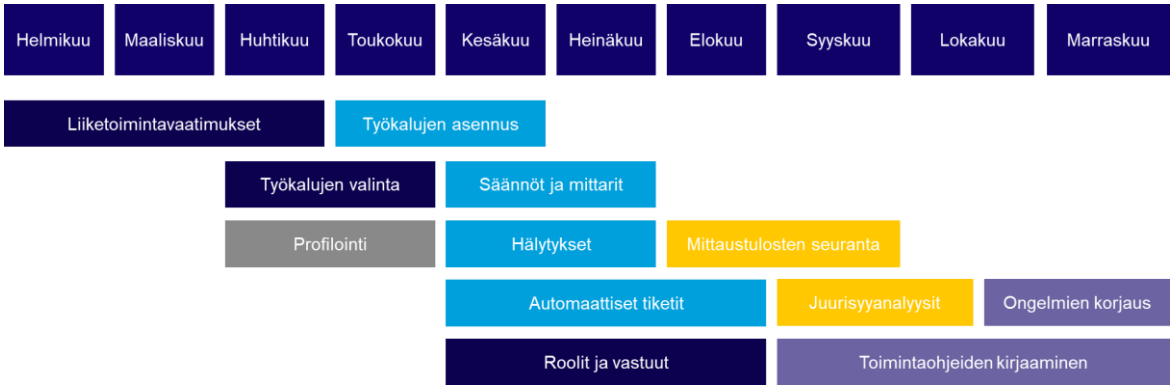
Kuva 10: Datan laadunhallinnan viitekehys

Viitekehys (kuvassa 10) on jaettu kirjallisuudessa esitettyyn viiteen pääkomponenttiin, jotka ovat määritä, mittaa, analysoi, kehitä ja ohjaa. Näiden komponenttien alle on jaettu erilaisia toimintoja ja aktiviteetteja, joita vaiheiden aikana tehdään. Katkoviivalla ympäröidyt toiminnot ovat sellaisia, jotka tyypillisesti tarvitsevat tehdä vain ensimmäisellä kerralla ja tarvittaessa päivittää. Seuraavassa kappaleessa syvennytään käyttöönoton vaiheisiin tämän projektin kontekstissa. Ohjaussykli jää tämän projektin ulkopuolelle aikataulurajoitteista johtuen.

5.1 Käyttöönoton vaiheistus

Datan laadunhallinnan viitekehyksen jalkauttamisen hanke lähti liikkeelle kuvan 11 mukaisesti helmikuussa, jolloin käytiin ensimmäiset keskustelut toimittajan kanssa.

Toimintatutkimuksen syklisen prosessin viittä vaihetta pyrittiin noudattamaan jokaisessa syklissä, mutta aivan kaikkien vaiheiden aikana esimerkiksi diagnosointia ei tehty, sillä suurin osa ongelman määrittelystä tehtiin projektin alkuvaiheessa. Profilointi jätettiin pois toteutuksesta teknisistä rajoitteista johtuen.



Kuva 11: Käyttöönoton vaiheistus

Ensimmäisessä syklissä (määritä) helmikuusta huhtikuuhun pohdittiin diplomityön rajausta, eli sopivaa pilotointikohdetta datan laadunhallinnan menetelmien jalkautukselle. Tässä vaiheessa myös kartoitettiin liiketoimintavaatimuksia sekä tarkoitukseen sopivia työkaluja. Muutamien läpikäyntien jälkeen ja tämän syklin arviointivaiheen jälkeen todettiin, että tekninen ratkaisu voisi olla Helenin ympäristöön sopiva, mutta sen käyttöönotto tehtäisiin sisäisesti omilla kyvykkyyksillä ja resursseilla. Maalis-huhtikuussa käytiin liiketoimintojen sekä teknisen tiimin kanssa läpi vaatimuksia laadunhallintamenetelmien osalta. Tähän kartoitukseen otettiin mukaan salkkuennusteen asiantuntijoita, datainsinööri sekä prosessin vastuuanalyytikko.

Toisen syklin (mittaa) aikana toukokuussa aloitettiin varsinainen mittausvaihe, jolloin DQ-moduulin asennukset saatiin data-alustalle tehtyä. Toimittajan konsultti suoritti asennukset alustalle, jonka jälkeen projektitiimin datainsinööri alkoi työstämään teknistä toteutusta. Kesäkuussa laadittiin teknisen arkkitehtuurin muista komponenteista suunnitelma, jonka perusteella tarvittavat DQ-moduulit ja muut palvelut saatiin otettua toimintaympäristössä käyttöön. Viikoittaisia statuspalavereita projektiryhmän kesken alettiin pitämään kesäkuussa, jotta pysyttiin perillä tapahtumista ja tehtävistä. Heinäkuun aikana datainsinööri sai luotua tarvittavat toiminnallisuudet teknisen arkkitehtuurin mukaisesti pystyyn alustalle, jotta

salkkuennusteprosessin datojen osalta osa tärkeimmistä mittareista saatiin pystytettyä data-alustalle.

Kolmas sykli, eli analysointivaihe käynnistyi elokuussa, jolloin datan laadusta kyettiin keräämään tietoa ja esiintyvistä virheistä tehtyä hälytyksiä, sekä luomaan ongelmatilanteista tikettejä, joita datatiimi pystyi tutkimaan ja selvittämään. Projekti myöhästyi hieman alkuperäisestä aikataulusta, sillä vaatimuksia alkoi tulla lisää, ennen kuin alkuperäisiä perustoiminnallisuuksia (taulukko 10) oli saatu vielä valmiiksi. Tästä syystä tyydyttiin osin yksinkertaisimpiin toteutuksiin tiettyjen mittarien, kuten esimerkiksi datan tuoreuden osalta. Tämä todettiin kolmannen syklin arviointivaiheessa elokuun aikana. Elokuussa tehtiin vielä hieman vaatimusmäärittelyä raporttien osalta, tarkemmin ottaen siis, mitä tietoja tekniseen ja liiketoiminnan näkymiin haluttiin ja kuinka ne visualisoitaisiin.

Kesäkuun aikana alettiin myös kartoittamaan datan omistajuusrakennetta uudelleen ja varattiin alkusyksylle dataomistajien yhteistyöverkosto-työpaja. Tämän tarkoituksena oli käydä datan omistajuusrakennetta läpi ja sovittaa sitä niin, että laadunhallintanäkökulmaa korostettiin tiettyjen roolien vastuualueissa. Elokuussa pidettiin ensimmäinen dataomistajien yhteistyöverkoston tapaaminen, joissa nämä asiat tuotiin esille. Varsinainen työpaja, jonka agendana oli kirkastaa rooleja sekä vastuita datan laadun osalta varattiin kuitenkin marraskuulle.

Neljäs sykli (kehitä) keskittyi pääsääntöisesti toiminnan kehittämiseen sekä toteutuksen arviointiin. Alkusyksyn aikana analysoitiin datan laatua ja tyypilliset virhetilanteet sekä niiden oikaisuohjeet kirjattiin ylös helpottamaan vastaavien tilanteiden ratkomista tulevaisuudessa. Syklin aikana todettiin, että tekninen toteutus vaatisi vielä vähän työstöä muutaman funktion osalta, joita tutkimuksen kirjoittaja lähti muokkaamaan.

Näiden lisäksi varattiin aika ajoin toistuvia palavereita, joissa käsitellään Datan hallinnan ajankohtaisia asioita, kuten datan hallintaan liittyviä työn alla olevia ja tulevia töitä, jotka vaikuttavat vastuuanalyytikoiden / heimojen tekemiseen. Myöhemmissä tilaisuuksissa on teema, kuten esimerkiksi datakatalogin käyttöönotto tai datan laadun kehitykset. Näissä tilaisuuksissa on vastuuanalyytikoilla myös mahdollisuus nostaa esiin datan hallintaa koskevia asioita ja ongelmia, joihin he toivoisivat datan hallinta ja kyvykkyydet -tiimiltä tukea. Tilaisuudet toimivat myös mahdollisuutena kasvattaa tietoisuutta datan laadusta sekä liiketoiminnoissa että Data ja tekoäly-yksikössä.

Syklien 2–4 keskeiset palaverit ja työpajat ovat kuvattu taulukkoon 10.

Taulukko 10: Vaihe 2: Viitekehyksen sovittaminen Heleniin, sekä sen jalkauttamisen ja tulosten arviointi

Nro	Osallistuja(t)	Menetelmä	Aihe	Tarkoitus	Pvm
1	Data ja tekoäly-yksikkö, toimittajan konsultit	Teams-työpaja	Datan laadun viitekehysehdotuksen läpikäynti	Läpikäynti teknisestä näkökulmasta	25.4.2022
2	Data ja tekoäly-yksikkö, sähkökaupan asiantuntija, toimittajan konsultit	Teams-työpaja	Datan laadun viitekehysehdotuksen läpikäynti	Läpikäynti liiketoiminnan näkökulmasta	29.4.2022
3	Data ja tekoäly-yksikkö, toimittajan konsultit	Teams-työpaja	Datan laatumoduulin asennus ja opastus	Datan laatumoduulin asentaminen ja käyttöön opastaminen	23.5.2022
4	Datavirrat-tiimin vetäjä	Teams-haastattelu	Profilointityökalun käyttöön opastaminen	Profilointityökalun käyttöön opastaminen	23.5.2022
5	Data ja tekoäly-yksikkö, konsultti	Teams-työpaja	Viitekehyksen tekninen arkkitehtuuri	Teknisen arkkitehtuurin työstöä	13.6.2022
6	Data ja tekoäly-yksikkö	Teams-läpikäynti	Teknisen raportin vaatimusmäärittelyä	Mitä asioita DQ-raportilta halutaan seurata	11.8.2022
7	Dataomistajat, Data ja tekoäly-yksikkö	Teams-työpaja	Dataomistajien yhteistyöverkosto-työpaja	Datan laatuun liittyvien roolien ja vastuiden kirkastaminen	23.8.2022
8	Datainsinöörit	Teams-palaveri	DQ moduulin teknisten toiminnallisuuksien läpikäynti	DQ-moduulin lisäkirjastojen käytön opastus	29.8.2022
9	Datainsinöörit	Teams-palaveri	Logic Apps-workflow kehitys	Devops tikettien ja kommenttien luonti rajapinnan avulla	22.9.2022
10	Datainsinöörit, konsultti	Teams-läpikäynti	Lisäfunktioiden käyttö DQ-moduulissa	Great Expectations funktioiden lisääminen moduuliin	5.10.2022

5.1.1 Määritä

Datan laadunhallinnan viitekehyksen jalkauttamisen hanke lähti liikkeelle helmikuussa, jolloin käytiin ensimmäiset keskustelut toimittajan kanssa. Helenillä oli jo aiemmin kokeiltu Great Expectations-python kirjastoa, johon toimittajankin moduuli pohjautui, jonka vuoksi työkalu nähtiin sopivan käyttötarkoitukseen. Tässä syklissä tutkimuksen kirjoittaja haastatteli sekä toimittajan konsultin kanssa että itsenäisesti sisäisiä asiantuntijoita tarpeiden kartoittamiseksi. Syklin lopuksi arvioitiin, että tekninen ratkaisu voisi olla Helenin ympäristöön sopiva, mutta sen käyttöönotto tehtäisiin sisäisesti omilla kyvykkyyksillä ja resursseilla.

Samoihin aikoihin myös profilointityökalua ja sen toimivuutta testattiin, sillä dataprofiileja ei ollut aiemmin saatavilla salkkuennustedatasta muutoin kuin kultatason tauluista. Huomattiin kuitenkin, ettei profilointityökalu toiminutkaan odotetulla tavalla, mikä hidasti tämän vaiheen toteutusta. Syy oli, että toisen toimittajan aiemmin syksyllä laatima python-automaatio profiilien generoimiselle oli toisesta kirjastosta versioriippuvainen. Ongelmaa ratkottiin datainsinöörien kesken ja tutkimuksen toteuttaja osallistui näihin palavereihin sekä selvitystyöhön myös. Lopulta päädyttiin aikataulu- ja resurssirajoitteista johtuen kuitenkin siihen lopputulokseen, ettei profilointitoiminnallisuutta tämän projektin puitteissa saada toteutettua, vaan odotetaan, että datakatalogiin tulee vastaava natiivi ominaisuus.

5.1.2 Mittaa

Toukokuussa tarvittavat asennukset saatiin valmiiksi data-alustalle, minkä jälkeen ensimmäisiä testauksia alettiin toteuttaa. Testauksen alkuvaiheessa päätettiin, että laitetaan ensisijaisesti pystyyn dataputkien terveysmittarit ja sen jälkeen pohditaan salkkuennustedatan sisältömittareita. Nykytila-analyysista ja liiketoimintavaatimusten kartoituksesta tunnistettuja datan laadun mittareita on esitetty taulukossa 11. Taulukossa on mittarin tyyppi (tekninen vai sisällöllinen), nimi, kuvaus sekä mihin datan laadun dimensioon sen voidaan katsoa kuuluvan.

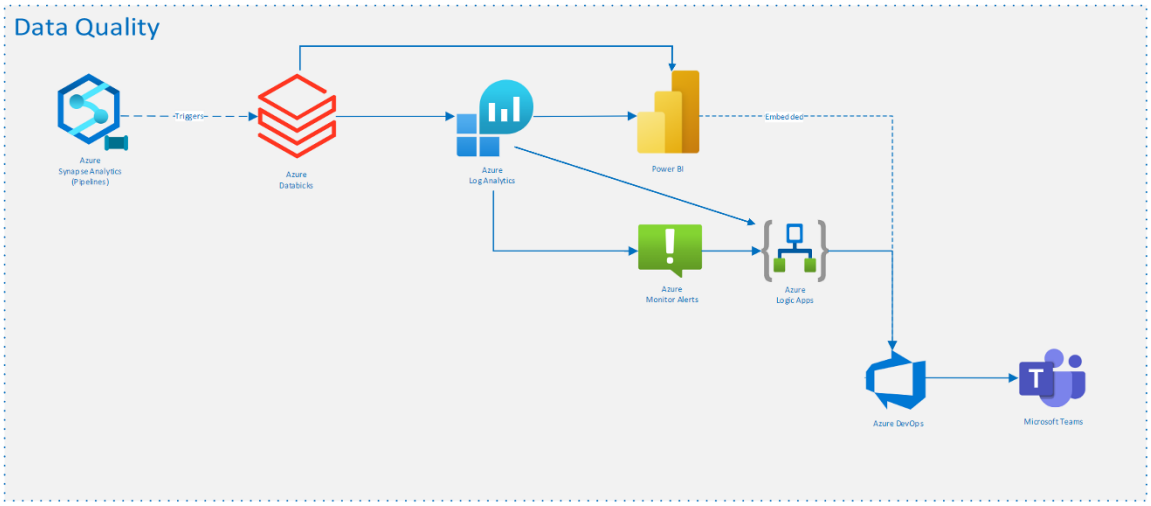
Taulukko 11: Datan laadun mittareita ja määritelmiä

Mittarin tyyppi	Nimi	Kuvaus	Laadun ulottuvuus
Tekninen	Rivimäärä	Tarkistetaan latausten rivimäärät, että tiedetään jos latauksissa on ongelmia (esim. päivittäin tulee x määrä rivejä)	Täydellisyys
Tekninen	Ajojen / latauksien kesto	Latauksen oletettu kesto tiedossa, että tiedetään jos latauksissa on ongelmia (oletettu kesto 1h, mutta nyt kestänyt 3h)	Ajantasaisuus
Tekninen	Tuoreus	Datan pitää olla vähintään yhtä tuoretta kuin x päivää/tuntia/tmv ennen nykyistä aikaleimaa (esim. päivän vanhaa, ei vanhempaa)	Ajantasaisuus
Tekninen	Skeeman validointi	Onko skeema muuttunut? Oletetaan että taulun sarakkeet ovat x, y ja z.	Oikeellisuus
Sisällöllinen	NULL-arvojen määrä	Puuttuuko oleellisia tietoja (nulleja yleensä 50 % arvoista, nyt 80 %)	Täydellisyys
Sisällöllinen	Uniikkius	Oletetaan, että kenttä saa ainoastaan uniikkeja arvoja, ei duplikaatteja	Ainutlaatuisuus
Sisällöllinen	Arvojen distribuutio	Kentän arvot pitää olla määritetyllä välillä tai saada määritetyn logiikan mukaisia kategorisia arvoja.	Oikeellisuus
Sisällöllinen	Oikeinkirjoituksen tarkistus	Kentän arvojen pitää noudattaa tiettyä kaavaa / olla tietyssä muodossa.	Oikeellisuus

Dataputkien terveysmittareiksi valittiin havaittavuuden periaatteita noudattaen tuoreus, volyymi, skeema ja arvojen distribuutioiden seuranta. Ajojen keston seuranta tapahtuu lokitietojen avulla. Datan elinkaaren visualisoinnin osalta kehitys on vielä kesken, sillä nykyistä putkista oikeanlaisen teknologian puutteen vuoksi ei ole mahdollista tuottaa tarpeeksi kattavaa elinkaaren visualisointia. Myös volyymin seuranta todettiin haasteeksi, sillä sääntömääritykset eivät vielä tue dynaamisia parametreja. Tuoreus todettiin näistä mittareista hankalimmaksi toteutuksen osalta, sillä siihen vaadittua funktiota ei ollut asennetussa dq-moduulissa valmiina, vaan se jouduttiin tekemään kustomoituna funktiona. Muut tarkastukset, kuten skeeman validointi, uniikkius sekä täydellisyys pystyttiin toteuttamaan asennetun

moduulin valmiilla funktioilla. Näiden lisäksi laadittiin yksinkertaisia oikeellisuustestejä, eli minkä kenttien tulee noudattaa tiettyä kaavaa tai mitkä kentät eivät saa saada tyhjiä arvoja. Näitä mittareita varten sparrailtiin salkkuennusteprosessin liiketoiminta-asiantuntijoita, sillä heillä oli kattava ymmärrys datan sisällöstä ja minkälaisia arvoja niiden tulee saada.

Tässä vaiheessa projektia laadittiin myös tekninen arkkitehtuurisuunnitelma, joka on esitettynä kuvassa 12.



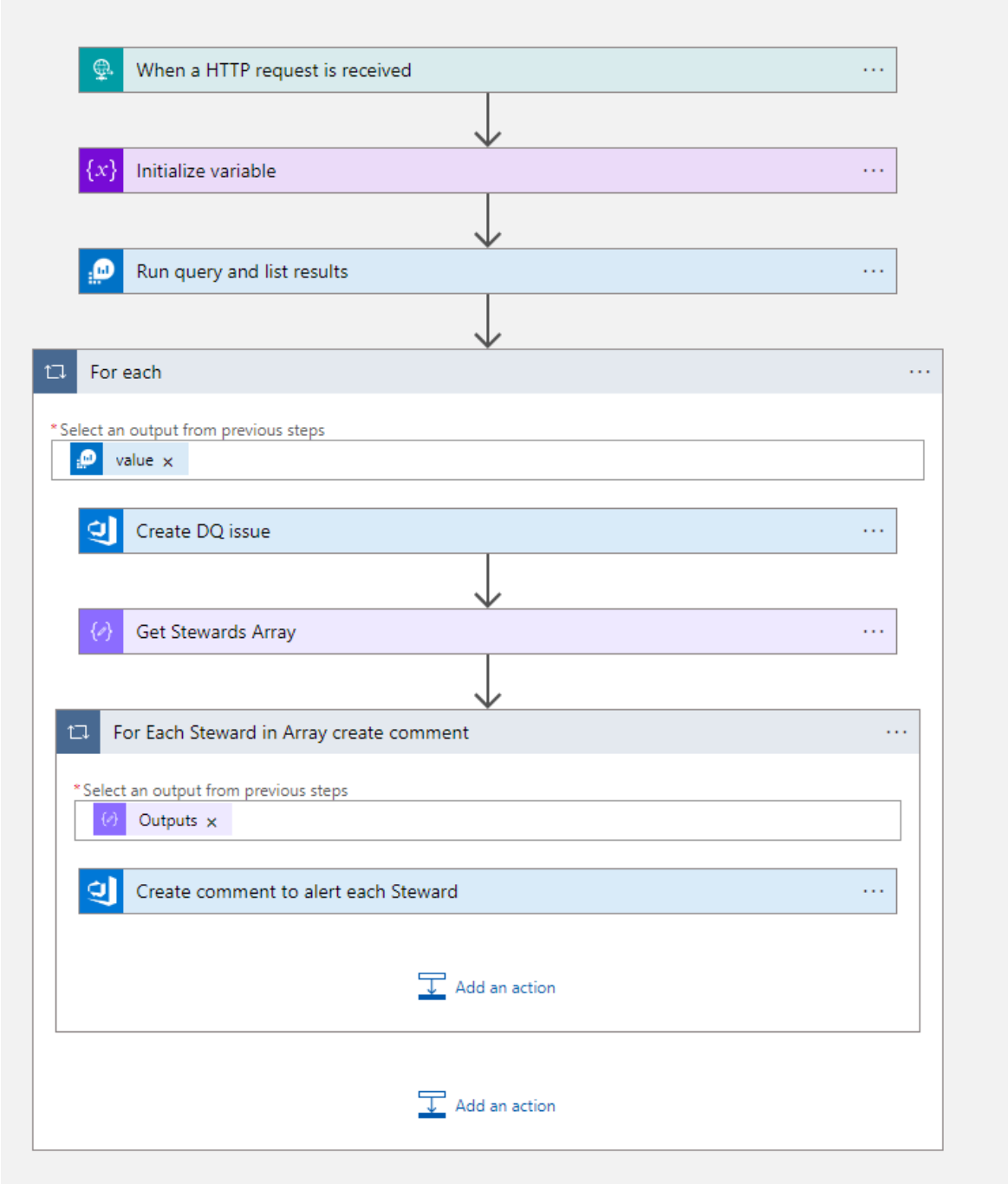
Kuva 12: Datan laadunhallinnan tekninen arkkitehtuuri

Säännöt muodostuvat Databricks työkirjojen sekä JSON-konfiguraatiotiedoston avulla. Datainsinööri kirjaa määritetyt säännöt JSON-konfiguraatiotiedostoon (kuvassa 13) sekä lisää Databricks työkirjakomponentin dataputken viimeiseksi vaiheeksi orkestrointityökalussa (Azure Synapse Analytics). Orkestrointityökalussa määritetään myös parametrit (taulu sekä mitaliarkkitehtuurin kerros) mihin säännöt halutaan kohdistaa, olettaen, että konfiguraatiotiedostot ovat luotu. Tämän seurauksena master-putkien ajon yhteydessä myös datan laadun säännöt ajetaan aina läpi ja testien tulokset tallentuvat deltatauluihin. Vaihtoehtoisesti säännöt voidaan myös ajaa manuaalisesti, mikäli data päivittyy tiheästi eikä testejä haluta tai ole tarpeellista ajaa liian usein.

```
1  [
2      {
3          "table": "table_name",
4          "ClDQRules": [
5              {"expect_column_values_to_be_unique": {
6                  "column": "column_name"
7                  }
8              }
9          ]
10     },
```

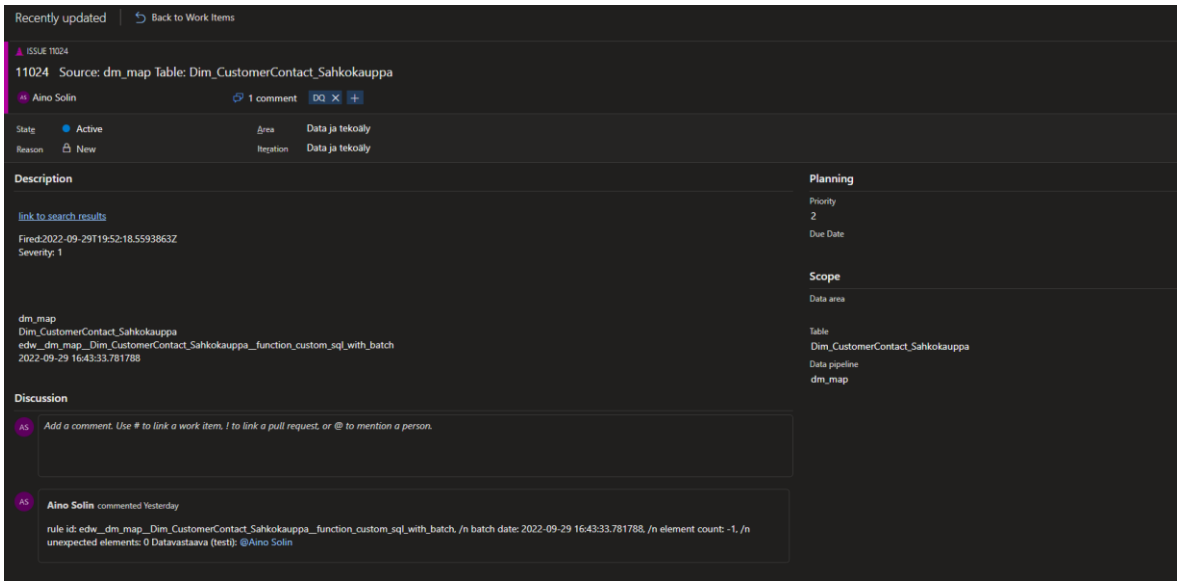
Kuva 13: JSON-konfiguraatiotiedosto

Hälytyksien osalta hyödynnettiin Log Analyticsia sekä Azure Monitor Alerts-työkalua, jonne voidaan parametrien avulla valikoida halutut seurattavat asiat. Deltatauluista luetaan tietyt sarakkeet, jotka kirjoitetaan Log Analyticsiin omaan tauluunsa. Monitor Alerts puolestaan seuraa tuota taulua ja onko sinne ilmestynyt rivejä, joissa validoinnin tulos on virheellinen. Logic Appsiin rakennettu logiikka (kuvassa 14) tunnistaa, mikäli Monitor Alertsista tulee hälytys, ja laatii tämän perusteella tiketin (work item) Azure Devops-työkaluun.



Kuva 14: Logic Apps-prosessi tikettien luontiin

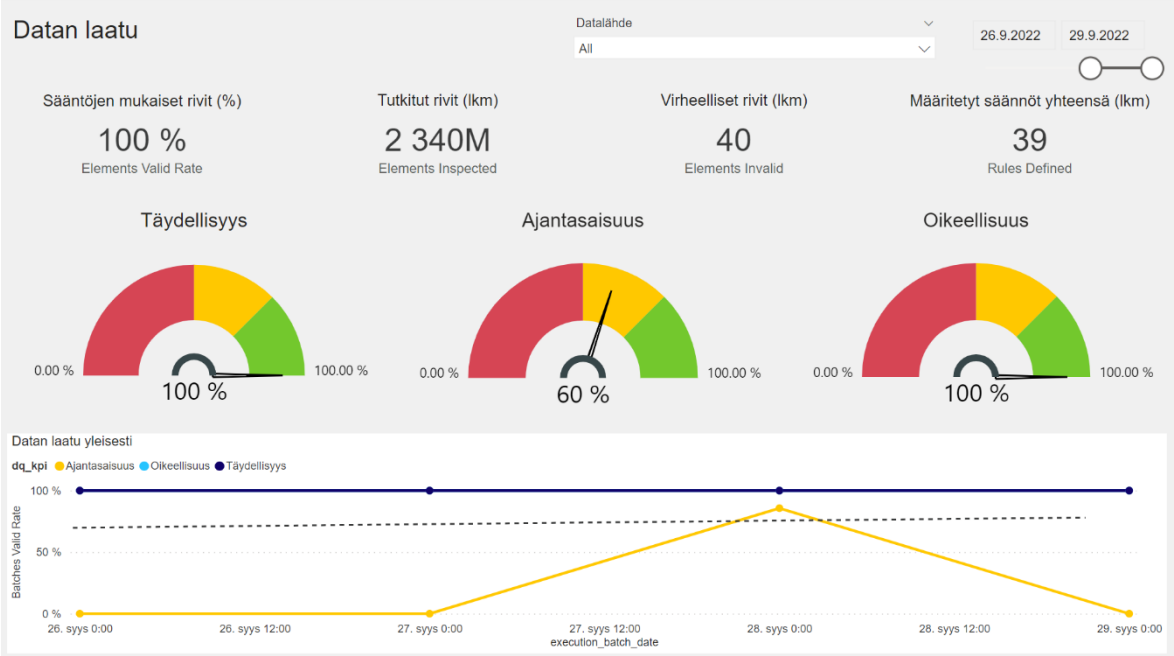
Tiketille (kuvassa 15) tuodaan tieto virheilmoituksesta ja kuvaukseen linkitetään virheen triggeröivä rivi, jota voi mennä tarkastelemaan Log Analyticsista. Lisäksi tikettiin kommentoidaan taulusta vastaava henkilö, jotta tämä saa myös sähköpostiinsa ilmoituksen, että datassa on havaittu virhe.



Kuva 15: DQ-tiketti Devopsissa (esimerkki)

5.1.3 Analysoi

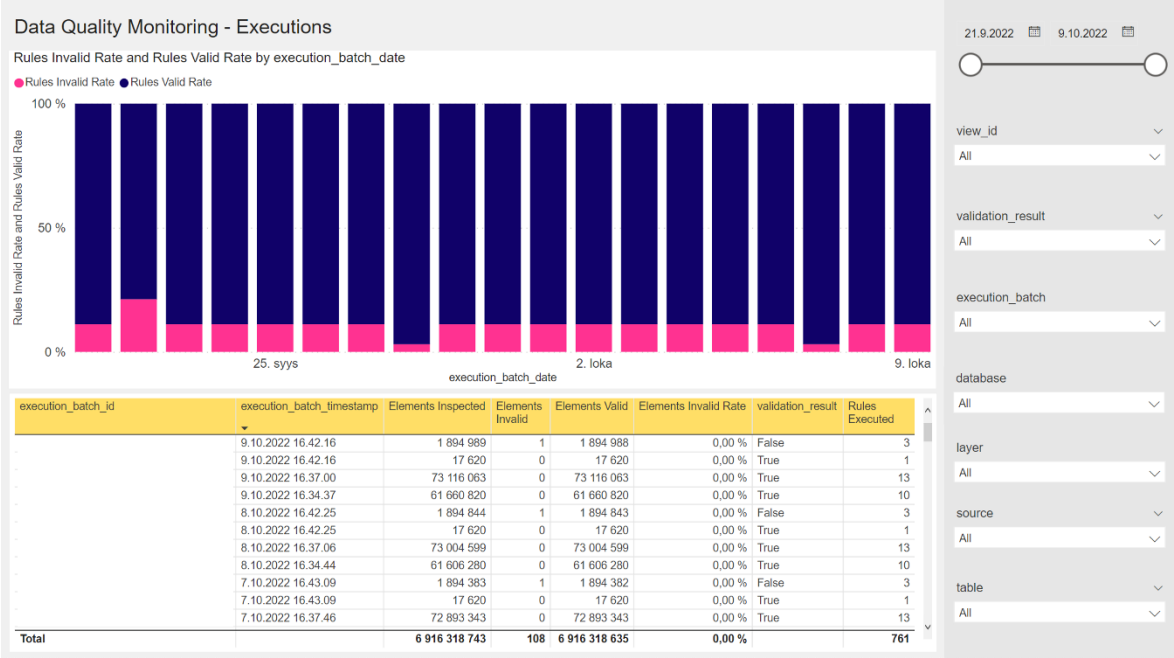
Tietojen analysointityökaluksi valikoitui PowerBI ja sinne rakennetut raportit, jonne kerättiin tietoa mainituista mittareista ja testien tuloksista. PowerBI-raportteja lähestyttiin karkeasti kahdesta näkökulmasta, joista toinen on pääsääntöisesti liiketoiminnan käyttöön sekä yleiseen datan laadun näkyvyyteen koko organisaatiolle, ja toinen virheitä selvittäville osapuolille, eli esimerkiksi Data ja tekoälytiimin datainsinööreille, analyytikoille sekä liiketoimintojen datavastaaville. Tärkeää oli laatia sekä liiketoimintaystävällinen näkymä, että teknisten ongelmien ratkaisuun suunnattu näkymä. Liiketoimintanäkymä pyrittiin rakentamaan riittävän yksiselitteiseksi ja kevyeksi, jota esimerkiksi ylempi johto tai muut liiketoiminta-asiantuntijat saavat tarvitsemansa tiedon tarpeeksi kompaktissa muodossa. Liiketoimintaraportti on esitetty kuvassa 16. Tätä raporttia voi hyödyntää sekä liiketoiminta, vastuuanalyytikko että muut prosessin sidosryhmät. Liiketoimintaraportille tuodaan pääsääntöisesti aggregoituja tuloksia, joilla pyritään antamaan kokonaiskuvaa siitä, missä kunnossa Helenin data-alustan data on.



Kuva 16: Liiketoiminnalle suunnattu datan laadun raportti (esimerkki)

Käyttäjä voi rajata raportilla näkyviä tietoja ajankohdan mukaan sekä lähdejärjestelmittäin. Tyypillisesti liiketoiminnoissa ollaan kiinnostuneita datan laadusta lähdejärjestelmätasolla, vaikka tietyn liiketoiminnan tietomallissa olisi yhdistelty eri lähteistä saatuja tietoja. Tämä antaa osviittaa siitä missä datavirhe on tapahtunut.

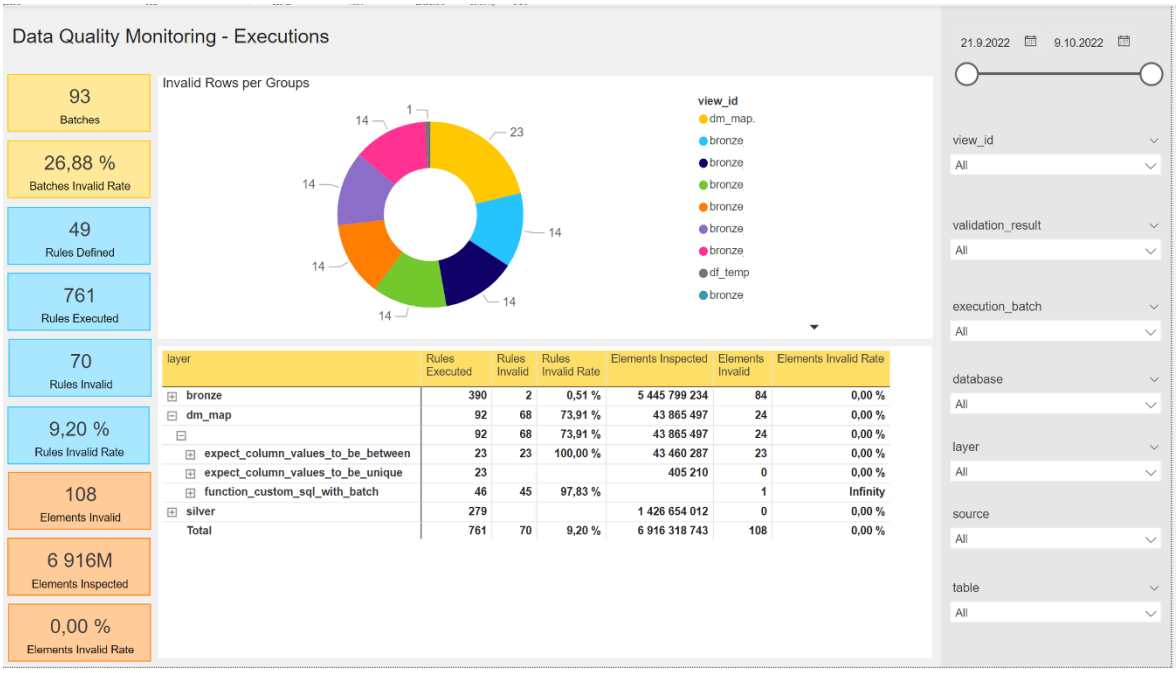
Tekninen raportti (kuvissa 17 ja 18) on pääsääntöisesti luotu virheiden selvitystyötä ja juurisyiden selvittämistä varten, sillä sinne tuodaan hienojakoisempaa tietoa, mihin käyttäjät voivat porautua vaikkapa rivitasolle saakka.



execution_batch_id	execution_batch_timestamp	Elements Inspected	Elements Invalid	Elements Valid	Elements Invalid Rate	validation_result	Rules Executed
	9.10.2022 16.42.16	1 894 989	1	1 894 988	0,00 %	False	3
	9.10.2022 16.42.16	17 620	0	17 620	0,00 %	True	1
	9.10.2022 16.37.00	73 116 063	0	73 116 063	0,00 %	True	13
	9.10.2022 16.34.37	61 660 820	0	61 660 820	0,00 %	True	10
	8.10.2022 16.42.25	1 894 844	1	1 894 843	0,00 %	False	3
	8.10.2022 16.42.25	17 620	0	17 620	0,00 %	True	1
	8.10.2022 16.37.06	73 004 599	0	73 004 599	0,00 %	True	13
	8.10.2022 16.34.44	61 606 280	0	61 606 280	0,00 %	True	10
	7.10.2022 16.43.09	1 894 383	1	1 894 382	0,00 %	False	3
	7.10.2022 16.43.09	17 620	0	17 620	0,00 %	True	1
	7.10.2022 16.37.46	72 893 343	0	72 893 343	0,00 %	True	13
Total		**6 916 318 743**	**108**	**6 916 318 635**	**0,00 %**		**761**

Kuva 17: Tekninen näkymä (1/2) (esimerkki)

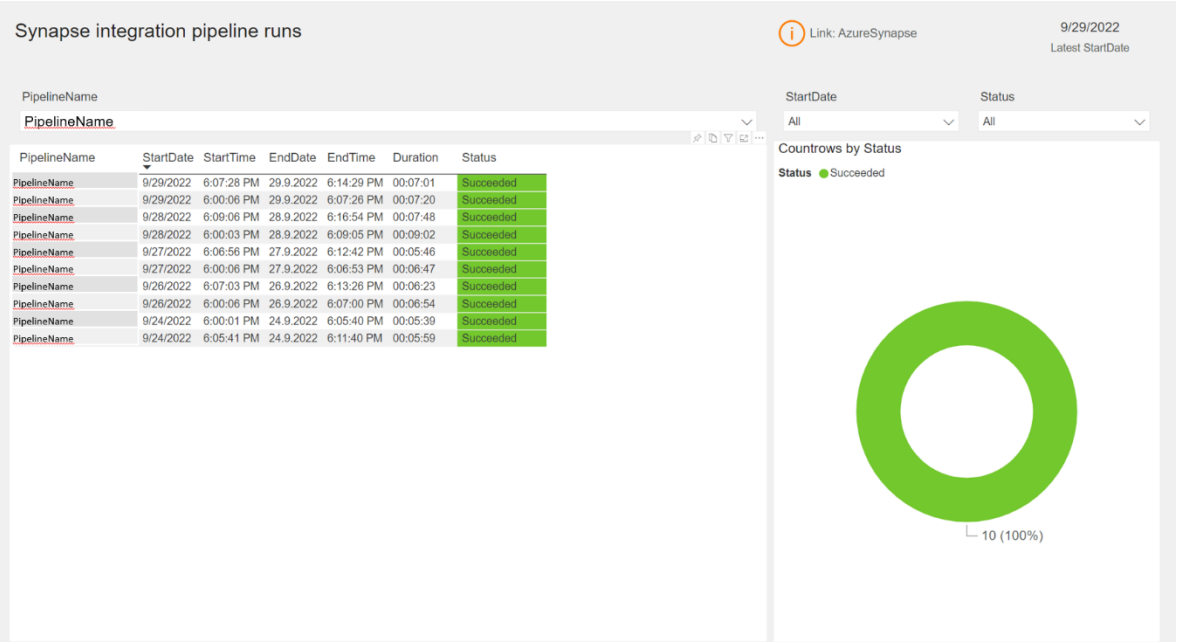
Käyttäjä voi myös rajata tietoja haluamaansa lähdejärjestelmään tai mitaliarkkitehtuurin tasolle (raaka/pronssi/hopea/kulta), voidakseen kartoittaa tarkemmin missä virhe on tapahtunut tai onko jokin sääntö tai taulu, joihin virheitä kohdistuu enemmän.



layer	Rules Executed	Rules Invalid	Rules Invalid Rate	Elements Inspected	Elements Invalid	Elements Invalid Rate
⊞ **bronze**	**390**	**2**	**0,51 %**	**5 445 799 234**	**84**	**0,00 %**
⊟ **dm_map**	**92**	**68**	**73,91 %**	**43 865 497**	**24**	**0,00 %**
⊟	**92**	**68**	**73,91 %**	**43 865 497**	**24**	**0,00 %**
⊞ **expect_column_values_to_be_between**	**23**	**23**	**100,00 %**	**43 460 287**	**23**	**0,00 %**
⊞ **expect_column_values_to_be_unique**	**23**			**405 210**	**0**	**0,00 %**
⊞ **function_custom_sql_with_batch**	**46**	**45**	**97,83 %**		**1**	**Infinity**
⊞ **silver**	**279**			**1 426 654 012**	**0**	**0,00 %**
Total	**761**	**70**	**9,20 %**	**6 916 318 743**	**108**	**0,00 %**

Kuva 18: Tekninen näkymä (2/2) (esimerkki)

PowerBI:hin luotiin sovellus, jonka alle tekniset näkymät tuotiin. Edellä mainittujen datan laadun raporttien lisäksi luotiin erillisiä näkymiä, jotka sisältävät tietoa dataputkien terveydestä, eli siis ajojen läpimenosta ja mihin ajankohtiin niitä on ajastettu. Yleisnäkymä ajojen läpimenosta on esitetty kuvassa 19 ja tuntinäkymä on esitetty kuvassa 20.



Kuva 19: Yleisnäkymä dataputkien terveydestä (esimerkki)

Datainsinööri voi valita raportilta haluamansa dataputken valikosta ja tarkastella aiempia läpimenoja eri ajankohtina tai statuksen mukaan, sekä siirtyä Azure Synapsen monitorointisivulle kyseisen ajon tarkempiin tietoihin klikkaamalla raportin oikeassa yläkulmassa olevaa oranssia infopainiketta.

Synapse integration pipeline runs

PipelineName: All

Status: All

Pvm	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	Total
9/29/2022	48	37	57	60	31	36	28	18	16	12	41	20	56	23	18	17	20	34	45	35	13				665
9/28/2022	34	44	56	51	31	34	27	16	20	11	40	15	50	17	11	13	18	29	61	34	21	36	25	27	721
9/27/2022	52	33	55	55	16	25	22	24	19	18	26	14	53	23	14	15	18	32	63	31	19	31	27	24	709
9/26/2022	1		2		5	24	24	15	17	16	40	14	45	19	10	10	15	27	53	37	19	32	27	25	477
9/25/2022	53	38	58	69	42	42	50	46	26	32	54	23	53	9					2			5		1	603
9/24/2022	6		2			3	18	30	32	26	50	25	60	28	25	23	29	42	65	34	26	42	29	31	626
9/23/2022	54	45	60	66	35	38	11										1					2		1	311
9/22/2022	60	39	62	66	43	45	30	29	29	24	51	24	45	29	22	24	26	40	62	44	30	45	33	33	933
9/21/2022	59	43	51	48	41	45	27	26	29	18	51	19	57	30	23	23	27	41	62	42	27	40	32	33	894
9/20/2022	75	62	83	73	57	57	43	30	26	24	58	20	66	25	23	23	25	36	61	40	29	49	34	39	1,057
9/19/2022	62	51	76	67	44	54	32	33	31	34	61	25	68	33	38	37	41	58	71	54	41	61	42	50	1,164
9/18/2022	63	48	82	72	45	50	42	39	36	29	59	32	71	30	25	25	30	43	66	41	29	46	30	38	1,069
9/17/2022	62	50	78	70	46	53	39	26	34	25	49	25	65	29	25	25	28	41	66	40	28	42	27	37	1,008
9/16/2022	2		2			4	21	30	30	28	52	25	57	31	25	24	27	40	56	40	33	45	28	38	646
9/15/2022	2		2			3	27	31	31	25	52	26	62	30	22	24	11							1	349
9/14/2022									2								1							1	4
9/13/2022	2		2			4	20	38	32	37	55	28	74	34	28	26	33	45	67	42	34	44	27	4	676
9/12/2022	61	55	74	77	50	40	13	2									1					4			379
9/11/2022	65	53	76	83	58	60	47	46	33	33	40	31	70	36	27	27	31	46	67	39	32	53	40	41	1,133
9/10/2022	2	1				2	2	23	35	27	55	28	68	33	27	27	31	43	67	43	30	49	33	37	663
9/9/2022	65	47	72	66	48	51	34	37	34	31	56	24	66	28	5		1					3			664
9/8/2022	56	50	85	69	48	52	37	31	36	28	58	32	72	36	26	26	35	45	71	49	39	45	34	39	1,079
9/7/2022	66	54	68	66	51	44	40	36	31	30	52	29	71	32	27	29	32	46	68	51	36	54	34	30	1,072
Total	1,366	1,079	1,540	1,484	991	1,071	867	810	765	677	1,307	654	1,615	767	596	585	685	1,032	1,506	1,015	739	1,129	762	844	23,886

Kuva 20: Kuormitus tunneittain (esimerkki)

Ajankohdat, joihin ajoja kohdistuu enemmän ovat merkitty tummalla sinisellä ja vaaleammalla merkittyinä aikoina ajojen kuorma on pienempi. Tämän matriisin avulla voidaan hahmottaa mitä putkia tulisi ajastaa putkia toisiin ajankohtiin, jotta kuormitus on tasaisempaa.

Näiden lisäksi haluttiin tietoa prosessiin kuluvista resursseista tai siitä, kuinka paljon virheitä esiintyy ja miten kauan niiden selvittämiseen kuluu aikaa. Tieto saadaan luettua suoraan Azure Devopsista natiivin PowerBI-datakonnektorin avulla. Tätä raporttia hyödynnetään lähinnä sisäisesti hahmottamaan, miten virheiden korjausprosessia voidaan parantaa sekä kuinka voitaisiin lisätä läpinäkyvyyttä. Liiketoimintaraportille on mahdollista tuoda jatkossa tästä datasta jalostettua metriikkaa. Resurssien seurannan raportin kehitys jäi vielä kesken, sillä muiden raporttien prioriteetti oli korkeampi.

5.1.4 Kehitä

Kehitysvaiheessa vielä työstettiin tiketöinnin teknisiä toiminnallisuuksia, sillä tieto datavastaavasta piti saada luettua konfiguraatiotiedostosta ja kirjoitettua se automaattisesti Logic Appsin avulla tiketin kommenttikenttään, jotta ilmoituksesta tulee myös sähköpostiviesti datavastaavalle. Tämä toiminnallisuus saatiin lopulta lokakuun aikana valmiiksi.

Jahka datan laadunhallintamenetelmiä ja työkaluja hyödynnetään käytännössä, voidaan alkaa kerryttää ymmärrystä yleisimmin toistuvista ongelmista ja toimintaa kehittää. Datatiimin

tulee vastuuanalyytikon ja datavastaavan johdolla nostaa esille laadussa esiintyvät puutteet heimokokouksissa ja yhdessä selvittää ongelman juurisyyt. Näin kasvatetaan myös liiketoimintojen puolella ymmärrystä datan laatutietoisuudesta. Selvitettyihin tiketteihin sen sulkemisen yhteydessä voidaan lisätä juurisyy, jonka avulla voidaan seurata yleisimmin esiintyviä virheitä ja jatkossa pyrkiä kitkemään nämä esimerkiksi kytkettäessä uusia datalähteitä data-alustalle.

Tähän projektiin yhtenä esimerkkinä otettiin sopimusdatassa esiintynyt virhe, jonka mukaan asiakkaan kulutustieto oli negatiivinen. Virhe havaittiin määritettyjen sääntöjen ja hälytysten avulla. Ongelmaa lähdettiin selvittämään aluksi kerros kerrokselta yksinkertaisella SQL-haulla, jonka jälkeen todettiin, että virhe oli tullut sellaisenaan data-alustalle. Tämä viittasi siihen, että virhe oli todennäköisesti tullut lähdejärjestelmästä tai virhe on johtunut esimerkiksi tiedonsiirrosta. Mikäli virhe on tapahtunut data-alustan ulkopuolella, niin datavastaavan tehtävä on selvittää ongelman syy tai eskaloida se seuraavalle virheestä vastaavalle taholle. Näin tässäkin tapauksessa toimittiin, eli kun todettiin ongelman tulleen sellaisenaan data-alustalle, tiketti eskaloitiin asiakaskohtaamisen tietoalueen datavastaavalle. Hänen vastuullansa oli katsoa, että ongelman syy selvitetään ja lopulta kirjata tiketti ratkotuksi.

Tämän esimerkkicasen sekä projektiryhmän kanssa keskustelun johdosta laadittiin joukko kysymyksiä ongelmanratkaisun tueksi:

Datan laadun sääntörikkomus havaitaan, tyypillisesti sääntö on määritetty yhdelle taululle tiettyyn kerrokseen:

- Missä virhe esiintyy ensimmäisen kerran? Tarkista raaka, pronssi, hopea, kulta, sekä raportointikerros. Mikäli ongelma on tullut sellaisenaan data-alustalle (eli esiintyy raakakerroksella), eskaloi datavastaavalle.

Ongelma ei todennäköisesti ole tullut sellaisenaan data-alustalle, vaan johtuu esimerkiksi transformointi- tai laskentalogiikasta:

- Kohdista virheen sijainti, eli kerros, kuten edellisessä vaiheessa. Käy läpi transformointi- ja laskentalogiikka kyseisen kerroksen ja sarakkeen osalta. Milloin koodia on päivitetty / mitä on muutettu?

- Onko tieto väärin kaikkien rivien osalta vai vain osittain? Koskeeko virheet vain tiettyä latausta / ajankohtaa? Onko skeema muuttunut (muutettu, lisätty tai poistettu sarakkeita)? Ovatko datatyypit tai formaatit ennallaan?
- Tarkista dataputkien tilanne, eli onko ajoja kaatunut / ovatko ne vielä kesken ym.
- Tarkista orkestrointityökalun virheloki, jos sellainen on saatavilla. Tämä todennäköisesti antaa virheen triggeröivän työkirjan ja rivin koodissa.
- Onko ympäristössä tapahtunut muutoksia, jotka olisivat voineet aiheuttaa virheitä järjestelmään? Onko annettu häiriötiedotteita tai tiedotettu huoltokatkosta? Tarkista myös ulkopuolisten palveluntarjoajien tilanne, eli Synapse, Databricks, ym. palvelujen tilanne.
- Onko prosessissa tapahtunut muutoksia, kuten latausten ajankohdan muuttuminen? Onko prosessissa esiintynyt vastaavaa ongelmaa aiemmin?
- Tarkista myös konfiguraatiotiedoston rakenne, onko JSON oikeassa muodossa?

Ongelma on raportointikerroksella:

- Lue data sisään esimerkiksi tyhjään raporttiin ilman käsittelyä ja vertaa virheelliseen dataan.
- Käy läpi transformointi- ja laskentalogiikka kyseisen virheen osalta yksi käsittelylogiikka kerrallaan ja yritä havaita milloin virhe esiintyy ensimmäisen kerran.

Kun ongelman syy on selvinnyt, ongelman korjausprosessi tulee käynnistää sekä kommunikoida seuraavat askeleet prosessin osallisille. Kommunikointikanava lähtökohtaisesti on Devops tiketin keskustelu tai Teams, mutta ongelman laajuudesta tai vakavuudesta riippuen saatetaan hyödyntää esimerkiksi Teams-kanavia ja sähköpostia. Ongelman korjaamisen jälkeen on hyvä dokumentoida syy ja ratkaisu sekä tiketille, että Teams-kanavan Wiki-sivulle.

Vastuista laadittiin myös alustava datan laadunhallinnan RACI malli yhdessä datan hallinta ja kyvykkyydet-tiimin voimin sekä dataomistajien tuella. Malli on esitetty kuvassa 21.

	Heimo	Sovellusvastaava	Datan hallinta ja kyvykkyydet	Datan laatuinsinööri	Datavastaava	Vastuuanalyytikko	Datainsinööri
Liiketoiminta-impaktin analysointi	CI		A	C	R	R	
Datan laadun tarpeiden kartoitus	C		A	C	I	R	CI
Datan laadun arviointi	I		A	C	I	R	CI
Liiketoiminnan sitouttaminen	I		AR	I	CI	CI	I
DQ-mittarien määritys, arviointi ja priorisointi	C	C	A	R	C	R	CI
Datan laadun tarkastelu ja monitorointi	I		A	R	R	R	I
Datan laatuvirheiden seuranta			A	R	R	R	R
Datan laatuvirheestä raportointi	I		CI	AR	R	R	R
Juurisyyanalyysi	CI	C	CI	AR	R	R	R
Ongelman korjaus	I	C	I	AR	R	R	R

Kuva 21: Datan laadunhallinnan roolit ja vastuut Helenillä

6 Johtopäätökset

Tässä luvussa tarkastellaan tutkimuksessa tehtyjen havaintojen ja tulosten perusteella vastauksia asetettuihin tutkimuskysymyksiin. Tutkimuksessa esille nousseet ylätason havainnot olivat usein tunnistettavissa teoriasta sekä aiemmasta tutkimuksesta, mutta tutkimuksen edetessä todettiin, että yksityiskohtaiset ratkaisut perustuivat enemmän asiantuntijahavaintoihin ja käytännön kokemukseen, sillä organisaatiossa käytettyihin teknologioihin tai haluttuun arkkitehtuuriin yhteensopivaa tutkimusta tai teoriaa löytyi hyvin vähän, jos lainkaan.

6.1 Datan laadunhallinnan käyttöönotto

Tämän alaluvun tarkoituksena on pyrkiä vastaamaan päätutkimuskysymykseen: *Miten datan laadunhallinnan menetelmät ja periaatteet voidaan ottaa käyttöön Data ja tekoäly -yksikön prosesseissa ja pienasiakasmyynnin salkkuennusteprosessissa?*

Datan laadunhallinnan pääperiaatteet voidaan tiivistää viiteen pääprosessiin, jotka ovat määritä, mittaa, analysoi, kehitä ja ohjaa. Nämä, kuten monet muutkin datan laadunhallinnan periaatteet perustuvat laatukirjallisuuteen. Viisi pääprosessia on lainattu DMAIC-mallista. Nämä viisi prosessia toimivat ohjenuorana siinä, miten laadunhallintamenetelmiä voidaan ottaa käyttöön organisaatiossa ja mitä asioita tulee ottaa huomioon. Malli on geneerinen, eli sitä voidaan soveltaa teknologiasta riippumatta, mutta jalkautus vaatii vaiheiden noudattamista ainakin ylätasolla ja menetelmien sovittamista kohdeorganisaatioon.

Määritä-vaiheessa pyritään kartoittamaan organisaation tarpeet ja vaatimukset koskien datan laatua sekä hahmottamaan, mitä organisaatiossa on tämän asian saralla mahdollisesti edistetty jo. Datan laadun osalta tämä tarkoittaa, että rakennetaan kattava ymmärrys datapääomasta, eli tunnistetaan, mikä on liiketoimintakriittistä tai korkean riskin dataa, jonka osalta laatua halutaan tai on kannattavaa seurata. Yrityksen tulisi keskittyä nimenomaan dataan, jonka heikolla laadulla voi olla merkittäviä vaikutuksia. Tässä vaiheessa keskitytään myös kuvaamaan tai keräämään olemassa oleva materiaali nykyisestä prosessista ja mahdolliset ongelmakohdat, joiden osalta asioita halutaan parantaa. Lisäksi hahmotetaan mitä käytössä olevia työkaluja voidaan hyödyntää datan laadunhallinnan arkkitehtuurissa ja rakennetaan

hahmotelma siitä. Salkkuennuste tunnistettiin liiketoimintakriittiseksi prosessiksi ja sen käyttämän datan laadussa oli jo aiemmin havaittu puutteita, jonka vuoksi sen laatua oli välttämätöntä alkaa seuraamaan. Huomattiin kuitenkin, että nykyisessä prosessissa on paljon manuaalisia- ja monivaiheisia tarkistuksia, joita jo alkuvaiheessa epäiltiin olevan liian monimutkaisia automatisoida tai tehdä koneellisesti kyseisillä työkaluilla. Tästä syystä päätettiin toteuttaa yksinkertaisemmat tarkistukset tämän projektin puitteissa ja monimutkaisemmat testit vasta kun työkalut ja menetelmät ovat otettu käyttöön.

Mittausvaiheessa olennaista on määrittää mitattavat asiat tarkemmin hyödyntäen aiemman vaiheen tuloksia. Liiketoimintakriittisen datan osalta valitaan esimerkiksi tietyt sarakkeet tai asiat, joita halutaan seurata. Tässä voidaan myös hyödyntää profiloinnin tuloksia, mikäli datan osalta ei välittömästi osata sanoa minkälaista kaavaa sen tulisi noudattaa, niin voitaisiin hyödyntää profiileja apuna datan laadun mittarien raja-arvojen määrityksessä. Laaja-alaisen datamassan profilointi koitui kuitenkin ongelmaksi, sillä datakatalogin taustalla oleva rajapinta rajoitti kutsuja, eli isoja määriä profiileja ei pystytty koneellisesti tuottamaan. Salkkuennusteen kultakerroksen datoista oli kuitenkin profiili olemassa, mutta niitä ei koettu tarpeellisiksi, sillä usein datan parissa työskentelevät analyytikot tai liiketoiminnan asiantuntijat osaavat kertoa miltä datan tulisi näyttää, kuten tässäkin tutkimuksessa. Varsinainen mittaus voidaan aloittaa, kun vaatimukset on tunnistettu ja työkalut mittausta varten pystytetty. Mittaus tapahtuu automaattisesti sääntöjen määrittämisen jälkeen, jahka ne on ajastettu esimerkiksi aina masterputken ajon yhteyteen, jolloin mittausdataa (salkkuennusteen datoista) saadaan päivittäin.

Datan laadun kunto konkretisoituu analysointivaiheessa, jonka aikana mittausvaiheen tuloksia päästään näkemään. Analysoinnin tarkoituksena on tuoda läpinäkyvyyttä datan laatuun, havaita virheet mahdollisimman aikaisessa vaiheessa ja tuloksien avulla tehdä johtopäätöksiä datan laadun virheiden synnystä. Tulokset tallentuivat datalaken deltatauluihin, joista arkkitehtuurin mukaisesti tieto ui sekä monitorointi- että raportointityökaluun. Tiedoista koostettiin datan laadun raportteja eri käyttäjäryhmille ja eri tarpeisiin. Liiketoiminnalle laadittiin yleiskuva datan laadusta ja sen kehityksestä eri dimensioineen (ajantasaisuus, täydellisyys ja oikeellisuus) ja käyttäjä voi rajata näkymän datalähteitä omien tarpeidensa mukaisesti. Juurisyyanalyysin tueksi laadittiin tekninen raportti, joka sisältää useita eri näkymiä, kuten dataputkien seurantanäkymän, ajojen kuormitus eri ajankohdittain sekä datan laadun mittarien yksityiskohtaisempi näkymä.

Kehittämisvaiheen prioriteettina on korjata analysoinnin aikana havaitut ongelmat ja viimeistään selvittää juurisyyt. Juurisyyanalyysissa voidaan jälleen lainata laatukirjallisuudesta tuttuja metodeja, kuten kalanruotokaaviota tai ”Five Whys”-metodia. Näiden avulla laadittiin yhtenä tuotoksena joukko kysymyksiä auttamaan virheen selvitystyössä. Kehittämisvaiheessa olennaista oli myös löytää sekä liiketoiminnalle että data- ja tekoäly-yksilölle sopivat yhteiset toimintamallit ja työkalut, jonka vuoksi päätettiin hyödyntää Azure Devopsia sekä Microsoft Teamsia tapaustenhallintatyökaluina. Suuri osa liiketoiminnasta sekä Data ja tekoäly-yksikkö käyttävät molempia työkaluja päivittäisessä työssään, jonka vuoksi ne koettiin tarpeeksi kevyeksi ratkaisuksi. Lisäksi tärkeää oli, että monitorointi tapahtuu pienellä vaivalla, eli se haluttiin automatisoida mahdollisimman pitkälle, sillä aiemmin virheiden seurantaan ja selvittelyyn kului paljon aikaa eikä vastuunjako ollut selkeää. Kehittämistyön tuloksena pystyttiin myös osittain jalkauttamaan datan omistajuusrakennetta, joka ei varsinaisesti kuulunut projektin rajaukseen, mutta joka nähtiin olennaiseksi osaksi datan laadunhallintaa. Tämän johdosta laadittiin RACI-malli datan laadunhallintaan.

6.2 Datan laadunhallinnan hyödyntäminen

Tämän alaluvun tarkoituksena on pyrkiä vastaamaan apututkimuskysymykseen: *Miten datan laadunhallintaa hyödynnetään?*

Datan laadunhallintaa hyödynnetään pääsääntöisesti saamaan arvoa datasta, josta jalostetaan tietoa. Mikäli päätöksenteossa käytetään virheellistä tai huonolaatuista tietoa, eivät päätöksenteon seurauksien voida olettaa olevan mieluisia, vaan pahimmassa tapauksessa ne voivat aiheuttaa tulonmenetyksiä. Teoria datan laadunhallinnasta esittää myös, että datan laadunhallinnan avulla voidaan lisätä läpinäkyvyyttä datan laatuun ja sitä kautta kasvattaa organisaation laatutietoisuutta. Mitä selkeämpää oman työn vaikutus datan laatuun on, sitä todennäköisemmin työn laatuun kiinnitetään enemmän huomiota ja virheiltä voidaan välttyä.

Datan laadunhallinta nähdään myös muutoksenhallinnan kriittisenä tukiprosessina, sillä monissa muutoksissa, kuten esimerkiksi tietojärjestelmäprojektien tai uuden kumppanuussuhteen alkaessa saatetaan luoda uusia rajapintoja tai yhdistää uusia tietolähteitä. Näissä yhteyksissä tyypillisesti asetetaan odotuksia ja vaatimuksia datan laadusta.

6.3 Datan laadun mittarit

Tämän alaluvun tarkoituksena on pyrkiä vastaamaan apututkimuskysymykseen: *Mitä datan laadun mittareita data-alustan ja salkkuennusteen osalta tulisi olla?*

Ennen tutkimusta varsinaisia datan laadun mittareita ei ole ollut eikä laadusta ole tuotettu raportointia yksikön ulkopuolelle, vaan ymmärrys datan laadun tasosta on ollut lähinnä Data ja tekoäly-yksikön sekä liiketoiminnan aktiivisimpien raporttien seuraajien tai datan hyödyntäjien tiedossa. Datan laadun seurantaan ei ollut myöskään yhdessä määritettyjä työkaluja tai standardisoitua tapaa mitata sitä.

Mittarit valikoituivat osittain teorian ja osittain organisaation asiantuntijoiden tietämyksen avulla. Teoriassa datan laatu jaetaan dimensioihin, joista osa on datalle luontaisia ja osa on järjestelmästä riippuvaisia. Tässä tutkimuksessa luontaisista ominaisuuksista puhutaan ”sisällöllisinä” mittareina ja järjestelmästä riippuvaisista ”teknisinä” mittareina. Sisällöllisten datan laadun mittarit saatiin pitkälti kartoitettua vastuuanalyytikon ja hinnoitteluasiantuntijan tietämyksellä. Mittareiksi valikoituivat salkkuennusteen datalähteiden osalta NULL-arvojen määrä (täydellisyys), duplikaatit (oikeellisuus), sekä arvojen distribuutio (oikeellisuus).

Datan sisällöllisten mittarien lisäksi on hyvä seurata dataputkien kuntoa, eli niin sanottuja teknisiä mittareita, sillä laskenta- ja transformointilogiikat voivat omalta osaltaan aiheuttaa datan laadun ongelmia. Teknisten mittarien valinnassa voidaan käyttää mallina havaittavuuden periaatteita, jotka ovat tuoreus, volyymi, skeema, jakauma ja datan elinkaari. Näistä valittiin seurattavaksi tuoreus (ajantasaisuus) ja skeema (oikeellisuus).

6.4 Tulosten ja luotettavuuden arviointi

Tutkimuksen tavoitteena oli jatkojalostaa laadunhallinnan viitekehystä Helenin toimintaympäristöön sopivaksi sekä jalkauttaa periaatteet ja työkalut data-alustalle sekä liiketoimintaprosessiin. Tämä kattoi salkkuennusteprosessissa käytetyn datan sisällön sekä dataputkien teknisten mittareiden määrittämisen, mittaamisen ja analysoimisen.

Määrittelyvaiheessa projektille asetettiin tiettyjä toiminnallisuusvaatimuksia, jotka ovat esitetty taulukossa 12. Tietyille toiminnallisuuksille annettiin korkeampi prioriteetti, sillä ne nähtiin välttämättöminä viitekehyksen jalkauttamisen onnistumisen kannalta.

Taulukko 12: Toivotut datan laadunhallinnan toiminnallisuudet tai konkreettiset tuotokset

Nro	Toiminnallisuus	Kuvaus	Prioriteetti	Otettu käyttöön
1	Dataputkien terveys (tekniset mittarit)	Tekniset mittarit, eli kuinka kauan ajot kestävät, kasaantuvatko ajot tietyille ajanjaksoille, ajot statuksien (succeeded, failed, jne.) mukaan.	1	
2	Datan sisällöllisset mittarit ja niiden tallennus	Datalle määritetään sisällöllisiä mittareita ja mittauksista syntyvät tulokset pystytään tallentamaan tietokantaan.	1	
3	Datan laadun raportointi	Mittareiden tuottama data visualisoidaan eri käyttötarpeisiin (liiketoiminta ja tekninen näkökulma).	1	
4	Hälytykset ja tiketöinti	Mittareille määritetään raja-arvot, joiden mukaan hälytyksiä ja tikettejä syntyy. Hälytykset Teams-viestinä, tiketöinti Azure Devopsiin.	2	
5	Datan elinkaari	Datan elinkaaren visualisointi (liikkuminen alkulähteeltä loppuraportille).	3	
6	Profilointi	Muiden taulujen profilointi eri mitaliarkkitehtuurin tasoilla (raaka, pronssi, hopea) olemassa olevalla työkalulla.	3	
7	Poikkeamien (anomaly) tunnistaminen	Poikkeamien tunnistaminen koneoppimisen avulla (Metrics Advisor).	3	
8	Datan laadunhallintaan kuluvien resurssien seuranta	Devops tiketteihin kuluva aika, paljonko auki / kiinni tikettejä, ym.	3	

Ne datan laadunhallinnan toiminnallisuudet, mitkä saatiin otettua käyttöön on merkitty taulukkoon vihreällä. Punaisella merkittyjä ei joko ehditty tai niitä ei ollut vielä mahdollista

toteuttaa käytössä olevilla työkaluilla. Kaikki ensimmäisen ja toisen prioriteetin toiminnallisuudet saatiin onnistuneesti otettua käyttöön, mutta kolmannen prioriteetin toiminnallisuuksista luovuttiin osittain teknisistä- ja aikataulullisista rajoitteista johtuen.

Datan elinkaaren visualisointi ei tällä hetkellä ole mahdollista käytössä olevan datakatalogin kautta, sillä se ei vielä tue Databricksia lähteenä, jossa suuri osa dataputkien transformointilogiikoista sijaitsee. Profilointia varten oli jo aiemmin laadittu joukko python-kirjastoja, mutta datakatalogin käyttämä rajapinta rajoittaa kutsuja, jonka vuoksi datakatalogin suurta datamassaa ei pystytty profiloimaan. Vaihtoehtoja profilointiin kuitenkin todettiin olevan muutamia, joista ensimmäinen olisi muokata kirjastoja käymään läpi ja profiloimaan datamassat niin sanotusti ”sivutetusti” rajapinnan antaman rajoituksen mukaisesti sen sijaan, että yritetään tehdä vastaava liian isolle määrälle dataa. Toinen vaihtoehto on odottaa, että käytössä olevaan datakatalogituotteeseen tulee natiivi profilointitoiminnallisuus. Poikkeamien tunnistus todettiin jo toisen syklin aikana jäävän pois, sillä muut toiminnallisuudet oli saatava ennen tätä valmiiksi. Datan laadunhallintaan kuluvien resurssien seurantaraporttia ei myöskään ehditty tämän projektin puitteissa toteuttamaan. Kolmannen prioriteetin toiminnallisuudet jäävät siis kaikki jatkokehitystoimenpiteiksi.

6.5 Jatkokehitysehdotukset

Diplomityön tuotoksena kehittynyttä ja käyttöönotettua viitekehystä ei implementoitu laajemmin kuin rajauksen kohteena olevan salkkuennustedatan osalta. Selkeästi merkittävin osa tutkimukseen käytetystä ajasta kului teknisten toiminnallisuuksien pystyttämiseen ja niiden toiminnan varmistamiseen. Jatkossa kuitenkin näitä perustoiminnallisuuksia ei tarvitse pystyttää uudelleen, vaan pelkät vaatimusmäärittelyt sääntöjä koskien riittää, sekä itse sääntöjen luonti moduuliin. Säännöt voivat toki kompleksisuudestaan riippuen vaatia monimutkaistakin kustomoitua kyselyä tai ne saatetaan pystyä tekemään jo olemassa olevilla sääntömäärityksillä. Tulevaisuudessa suositellaan moduulin käyttöönottoa myös muiden liiketoimintaprosessien datojen osalta. Näiden pohjaksi olisi hyvä tehdä kattavampi liiketoimintavaatimusten kartoitus, eli pohtia mitä liiketoimintakriittinen data Helenillä on, eli esimerkiksi priorisoida datasetit tai niiden tärkeimmät sarakkeet. Jatkossa olisi myös hyvä tutkia, kuinka monimutkaisemmat salkkuennusteen laatutarkistukset voitaisiin automatisoida

mahdollisimman pitkälle, sillä niiden tarkistamiseen menee suhteellisen kauan aikaa, ja käsin tehtävät tarkistukset ovat aina virheille alttiita.

Teknisten läpikäyntien aikana todettiin, että tietyissä datan laadun tarkistuksissa, kuten latausten rivimäärien validoinnissa, voisi hyödyntää koneoppimista anomalioiden havainnointiin. Aikataulurajoitteista johtuen tätä toteutusta ei ehditty tämän tutkimuksen aikana tekemään, mutta valikoitiin kuitenkin Azuren Metrics Advisor tuohon tarkoitukseen sopivaksi työkaluksi. Tätä suositellaan yhtenä jatkotoimenpiteenä.

Dynaamisia parametreja vaativiin tarkistuksiin, kuten eri mitaliarkkitehtuurien välisiin rivimäärätarkistuksiin, olisi hyvä keksiä toteutustapa nykyisellä moduulilla. Yhtenä ratkaisuna tähän voisi olla evaluaatioparametrit. Great Expectations kirjasto (johon DQ-moduuli pohjautuu) tukee evaluaatioparametreja, eli parametrit voidaan tallentaa omaan paikkaan, josta ne haetaan aina sääntöjen validointivaiheessa. DQ-moduuli on kuitenkin hieman eri tavalla konfiguroitu, kuin Great Expectationsin omien dokumentaatioiden mukaisesti ja moduulia ajetaan jaetussa ympäristössä, jolloin tulisi ensin tutkia onko evaluaatioparametrien käyttö näillä konfiguraatioilla mahdollista.

Datan laadunhallinnan kulmakiviä ovat datan elinkaaren visualisointi sekä profilointi, mutta näitä kumpaakaan toiminnallisuutta ei saatu projektin aikana otettua käyttöön työkalurajoitteiden vuoksi. Helenillä käytössä olevaan datakatalogiin on kuitenkin tulevaisuudessa tulossa molemmat toiminnallisuudet ja näiden käyttöönottoa suositellaan, jahka se on mahdollista.

Tulevaisuudessa datan laadun merkitys tulee lisääntymään, jonka johdosta akateemisia tutkimuksia aiheesta tullaan näkemään entistä enemmän. Varsinaisia käyttöönottotutkimuksia olisi hyvä toteuttaa lisää, jotta ymmärrystä teoreettisten viitekehysten toimivuudesta voitaisiin kerryttää sekä nähdä, onko teorian ja käytännön välillä todellisuudessa suurta kuilua. Tekoälyn ja koneoppimisen hyödyntäminen datan laadunhallinnassa on alati kasvava aihe, josta kyllä tiettyjen tarpeiden osalta nähdään jo tällä hetkellä toteutuksia, mutta sinne suuntaan kehitys on menossa ja paljon on vielä tutkimatta.

7 Yhteenveto

Tutkimus tehtiin Helsingin kaupungin omistamaan energia-alan yhtiöön Heleniin, joka tarjoaa asiakkailleen sähköä, kaukolämpöä ja -jäähdytystä sekä erilaisia palveluja muun muassa energian pientuotantoon ja energiankäyttöön sekä sen tehostamiseen. Konsernissa työskenteli tutkimuksen kirjoitushetkellä noin 1000 osaajaa ja asiantuntijaa, tehden siitä yhden Suomen suurimmista energiakonserneista. Helen tuottaa energiaa Helsingissä sijaitsevilla voimalaitoksilla, tuotantolaitoksilla sekä yhtiön omistamien voimaosuuksien kautta. Sähkömarkkinoiden muutokset ovat vaikuttaneet vahvasti sähkön hintaan, minkä vuoksi myös Helenin tulee kiinnittää enemmän huomiota siihen, ovatko sähkön hinnanmuodostuksen taustalla olevat ennusteet tai niiden käyttämä data validia. Sähkön hinta on erittäin volatiili, eli se voi muuttua radikaalisti, kuten tutkimuksen kirjoitushetken ajankohtana vuonna 2022 on huomattu. Rajujen hinnanmuutosten aiheuttamaa tulonmenetysriskiä vähentääkseen Helen pyrkiikin suojaamaan kodeille ja pienyrityksille myytävän sähkön hinnat salkkuennusteen avulla. Suojaustaso määritetään kulutusennusteiden mukaan, joten ennusteiden käyttämän datan laatua tulee seurata.

Tutkimuksen kirjoittaja osana Data ja tekoäly-yksikköä toteutti nykytila-analyysin datan laadusta ja sen hallinnasta vuonna 2021, jonka tuloksien perusteella havaittiin puutteita datan laadussa ja sen hallinnassa. Analyysin perusteella laadittiin myös kuva tavoitetilasta, mitä lähdettiin tämän tutkimuksen tavoitteena toteuttamaan. Tutkimuksen tavoitteeksi asetettiin datan laadunhallinnan viitekehyksen jalostaminen ja menetelmien käyttöönotto data-alustalla sekä salkkuennusteprosessissa. Viitekehyksen pohjana käytettiin valitun toteutuskumppanin laatimaa viitekehystä, jota sovitettiin Helenin toimintaan sopivaksi. Tutkimuksessa kartoitettiin myös liiketoimintavaatimuksia ja haluttuja toiminnallisuuksia esimerkiksi raportoinnin ja teknisten toteutusten osalta, sillä nykytila-analyysissa nousseet asiat olivat melko ylätasolla. Menetelmien käyttöönotto kattoi niiden konkreettisten teknologisten toiminnallisuuksien pystyttämisen data-alustalle, joiden avulla datan laatua voidaan mitata, monitoroida ja analysoida sekä datan laadunhallinnan yhteistyömallien ja toimintatapojen kehittämistä Data ja tekoäly-yksikön ja liiketoiminnan välille.

Tutkimuksen tuotoksena saatiin tarvittavat perustoiminnallisuudet pystyyn datan laadun mittaamisen, monitoroinnin ja analysoinnin osalta. Näiden lisäksi kehitettiin vastuumatriisi

datan laadunhallintaan sekä juurisyyn selvittämisen tueksi joukko apukysymyksiä. Kokonaisuudessaan tutkimuksen voidaan nähdä onnistuneen ja saavuttaneensa asetetut tavoitteet. Työn lopussa tuodaan myös esille työn rajoituksia sekä jatkokehitystoimenpiteitä, liittyen muun muassa siihen, että osa halutuista toiminnallisuuksista jäivät kesken tai niitä ei voitu toteuttaa teknisistä tai aikataulullisista rajoitteista johtuen.

Lähteet

Baskarada, S. ja Koronios, A. 2013. Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension. *Australasian Journal of Information Systems*, 18(1), s. 3–4. doi:10.3127/ajis.v18i1.748.

Batini, C. ja Scannapieco, M. 2016. Data and Information Quality Dimensions, Principles and Techniques. Cham Springer International Publishing, s. 23–28.

Batini, C., Rula, A., Scannapieco, M. ja Viscusi, G. 2015. From Data Quality to Big Data Quality. *Journal of Database Management*, 26(1). s. 3. doi:10.4018/jdm.2015010103.

Cai, L. ja Zhu, Y. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*. 14(0). doi:10.5334/dsj-2015-002.

Cichy, C. ja Rass, S. 2019. An Overview of Data Quality Frameworks. *IEEE Access*, 7, s. 24634–24648. doi:10.1109/access.2019.2899751.

D'Urso, C. 2016. Experience: Glitches in Databases, How to Ensure Data Quality by Outlier Detection Techniques. *Journal of Data and Information Quality*, 7(3), s. 1–22. doi:10.1145/2950109.

Data Management Association (DAMA) (n.d.). DAMA UK - Glossary. [verkkoaineisto] www.dama-uk.org. Saatavissa: https://www.dama-uk.org/Glossary [Viitattu 10.3.2022].

English, L. 2009. The TIQM ® Quality System for Total Information Quality Management: Business Excellence through Information Excellence. [verkkoaineisto] Saatavissa: http://mitiq.mit.edu.

Finto 2018. Tietämys. [verkkoaineisto] Finto.fi. Saatavissa: https://finto.fi/tt/fi/page/t37 [Viitattu 10.3.2022].

Flores, J. ja Sun, J. 2018. Information Quality Awareness and Information Quality Practice. *Journal of Data and Information Quality*, 10(1). s. 1–18. doi:10.1145/3182182.

Francisco, M.M.C., Alves-Souza, S.N., Campos, E.G.L. ja De Souza, L.S. 2017. Total Data Quality Management and Total Information Quality Management Applied to Costumer Relationship Management. *Proceedings of the 9th International Conference on Information Management and Engineering - ICIME 2017*. s. 3. doi:10.1145/3149572.3149575.

Ge, M. ja Helfert, M. 2013. 'Cost and Value Management for Data Quality', Sadiq, S. Handbook of data quality : research and practice. [E-kirja] s. 78-79. Heidelberg: Springer. Saatavissa: https://link.springer.com

Hassenstein, M.J. ja Vanella, P. 2022. Data Quality—Concepts and Problems. *Encyclopedia*. 2(1), s. 498–510. doi:10.3390/encyclopedia2010032.

Helen Oy. 2020a. Strategia | Helen. [verkkosivu] www.helen.fi. Saatavissa: https://www.helen.fi/helen-oy/helen-oy/tietoa-meista/strategia [Viitattu 7.11.2022].

Helen Oy. 2020b. Datastrategia [sisäinen dokumentti] [Viitattu 7.11.2022].

Helen Oy. 2022a. Digitaaliset Ratkaisut [intranet] [Viitattu 7.11.2022].

Helen Oy. 2022b. Energiakauppa - Salkunhoito [intranet] [Viitattu 7.11.2022].

International Organisation of Standardization 2008. ISO/IEC 25012:2008(en) Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. [verkkoaineisto] ISO Verkkoaineisto Browsing Platform. Saatavissa: https://www.iso.org/obp/ui/#iso:std:iso-iec:25012:ed-1:v1:en [Viitattu 14.3.2022].

International Organisation of Standardization. 2015. ISO 9000:2015(en) Quality management systems — Fundamentals and vocabulary. [verkkoaineisto] ISO Verkkoaineisto Browsing Platform. Saatavissa: https://www.iso.org/obp/ui/#iso:std:iso:9000:ed-4:v1:en:term:3.6.2 [Viitattu 8.3.2022].

ISO/IEC. 2020. ISO/IEC 25012:2020 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model. [verkkoaineisto] Saatavissa: https://sales.sfs.fi [Viitattu 14.3.2022].

Juuti, P. ja Puusa, A. 2020. Laadullisen tutkimuksen näkökulmat ja menetelmät. [verkkoaineisto] Gaudeamus. Saatavissa: https://www.ellibslibrary.com

Kaushik, S. 2020. Understanding the difference between Data Accuracy and Validity. LinkedIn. Saatavissa: https://www.linkedin.com/pulse/understanding-difference-between-data-accuracy-validity-kaushik/ [Viitattu 25.3.2022].

King, T. ja Schwarzenbach, J. 2020. Managing data quality : a practical guide. [verkkoaineisto] London: Bcs, s. 57–74. Saatavissa: https://search.ebscohost.com/

Kiran, D.R. 2019. Total Quality Management : key concepts and case studies. *Amsterdam I Pozostałe: Butterworth-Heinemann Is An Imprint Of Elsevier*. s. 1–4. doi:10.1016/C2016-0-00426-6

Lee, Y.W. ja Strong, D.M. 2003. Knowing-Why About Data Processes and Data Quality. *Journal of Management Information Systems*, 20(3). s. 13–39. doi:10.1080/07421222.2003.11045775.

Lee, Y.W., Pipino, L.L., Wang, R.Y. ja Funk, J.D. 2006. Journey to data quality. Cambridge, Mass.: Mit Press, s. 55–84.

Loshin, D. 2011. The practitioner's guide to data quality improvement. Burlington, Ma: Morgan Kaufmann, s. 1–406.

Madnick, S.E., Wang, R.Y., Lee, Y.W. ja Zhu, H. 2009. Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality*, 1(1). s. 1–22. doi:10.1145/1515693.1516680.

Mahanti, R. 2019. Data quality : dimensions, measurement, strategy, management, and governance. Milwaukee, Wisconsin: Asq Quality Press, s. 1–456.

Mahanti, R. 2021. Data Governance And Compliance : evolving to our current high stakes environment. S.L.: Springer Verlag, Singapor, s. 126.

McGilvray, D. 2008. Executing data quality projects : ten steps to quality data and trusted information. 1st ed. Academic Press, s. 31–32.

McGilvray, D. 2013. ‘Data Quality Projects and Programs’, Sadiq, S. Handbook of data quality : research and practice. [E-kirja] s. 41. Heidelberg: Springer. Saatavissa: https://link.springer.com/

MIT Critical Data. 2016. Secondary Analysis of Electronic Health Records. *Cham: Springer International Publishing*. s. 163-165. doi:10.1007/978-3-319-43742-2.

Moses, B. 2022. Data Quality Fundamentals : a practitioner’s guide to building trustworthy data pipelines. [E-kirja] S.L.: O’Reilly Media. Saatavissa: https://learning.oreilly.com.

Mosley, M., Brackett, M.H., Earley, S., Henderson, D. ja Data Administration Management Association 2010. The DAMA guide to the data management body of knowledge : (DAMA-DMBOK Guide). 1st ed. Bradley Beach, New Jersey: Technics Publications.

Olson, J.E. 2008. Data quality : the accuracy dimension. Amsterdam Morgan Kaufmann, s. 24–33.

Redman, T.C. 1998. The impact of poor data quality on the typical enterprise. *Communications of the ACM,* 41(2). s. 79–82. doi:10.1145/269012.269025.

Redman, T.C. 2013. ‘Cost and Value Management for Data Quality’, Sadiq, S. Handbook of data quality : research and practice. [E-kirja] s. 16, 26. Heidelberg: Springer. Saatavissa: https://link.springer.com.

Refsnes Data (n.d.). SQL Server ISDATE() Function. [verkkoaineisto] www.w3schools.com. Saatavissa: https://www.w3schools.com/sql/func_sql-server_isdate.asp [Viitattu 22.3.2022].

Sebastian-Coleman, L. 2013. Measuring data quality for ongoing improvement : a data quality assessment framework. Waltham, Ma: Elsevier.

Suojanen, U. 2014. Ulla Suojanen: Toimintatutkimus ammatillisen kehittymisen välineenä. [verkkoaineisto] METODIX. Saatavissa: https://metodix.fi/ [Viitattu 16.6.2022].

Taleb, I., Serhani, M.A. ja Dssouli, R. 2018. Big Data Quality: A Survey. *2018 IEEE International Congress on Big Data (BigData Congress)*. s. 4. doi:10.1109/bigdatacongress.2018.00029.

Tietoarkisto (n.d.). Toimintatutkimus. [verkkoaineisto] Tietoarkisto. Saatavissa: https://www.fsd.tuni.fi/fi/palvelut/menetelmaopetus/kvali/tutkimusasetelma/toimintatutkimus/ [Viitattu 16.6.2022].

Treder, M. 2020. Chief Data Officer Management Handbook : set up and run an organizations data supply chain. S.L.: Apress, s. 186–190.

Tuomi, J. ja Sarajärvi A. 2019. Laadullinen tutkimus ja sisällönanalyysi. Tammi, s. 88.

Valli, R. 2019. Ikkunoita tutkimusmetodeihin 1 : Metodin valinta ja aineistonkeruu: virikkeitä aloittelevalle tutkijalle. [verkkoaineisto] Ps-Kustannus. Saatavissa: https://www.elibslibrary.com. [Viitattu 23.6.2022]

Wang, R.Y. 1998. A product perspective on total data quality management. *Communications of the ACM,* 41(2). s. 58–65. doi:10.1145/269012.269022.

Wang, R.Y. ja Strong, D.M. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems,* 12(4). s. 5–33. doi:10.1080/07421222.1996.11518099.

Willis, J.W. ja Edwards, C.L. 2014. Action research : models, methods, and examples. Charlotte, Car. Du N.: Information Age Publishing. s. 13-60

Wook, M., Hasbullah, N.A., Zainudin, N., Jabar, Z., Ramli, S., Razali, N. ja Yusop, N. 2021. Big Data Quality Dimensions: A Systematic Literature Review. *Journal of Information Systems and Technology Management, 17*. s. 2–7. doi:10.4301/s1807-1775202017003.

Zhang, G. 2020. A data traceability method to improve data quality in a big data environment. *2020 IEEE Fifth International Conference on Data Science in Cyberspace (DSC).* s. 1. doi:10.1109/dsc50466.2020.00051.

Zhu, H., Madnick, S., Lee, Y. ja Wang, R. 2014. Data and Information Quality Research. *Computing Handbook, Third Edition.* s. 16–116–20. doi:10.1201/b16768-20.

Zuber-Skerritt, O. ja Wood, L. 2019. Action learning and action research : genres and approaches. Bingley, Uk: Emerald Publishing, s. 5

Liite 1: RACI-vastuunjakotaulukko datan laadunhallintaan

Taulukko 13: RACI-vastuunjakotaulukko datan laadunhallintaan (mukaillen Loshin 2011, s. 87-88).

	Liiketoi-mintajohto	Liiketoi-minta	Sovellus-vastaava	Datan hal-linnan joh-taja	Datan laadun johtaja	Datavas-taava	Data-analyy-tikko	Järjestelmä-kehittäjä / Datainsinööri	Operatiivinen henkilöstö
Liiketoimintavaikutuk-sen arviointi	A	CI			C		R		CI
Datan laadun tarpeiden kartoitus		A	CI	R	C		C	CI	
Datan laadun arviointi		I	CI	I	A	I	R	CI	CI
Liiketoiminnan sitout-taminen		CI	CI	A	R	C	C	CI	
DQ-mittarien määritys, arviointi ja priorisointi		A	CI	R	C	C	C	CI	CI
DQ-mittarien määritys	A	CI	CI		R	C	C		
Hyväksyntäkriteerien raja-arvojen määritys	A	A	CI	CI	R	C	C		
Datan validiteettisään-töjen määritys		A	CI	I	R	C	C		
Datan laadun tarkas-telu ja monitorointi		I	I	A	I	R			

(jatkuu)

Liite 1: RACI-vastuunjakotaulukko datan laadunhallintaan

Taulukko 6: RACI-vastuunjakotaulukko datan laadunhallintaan (jatkuu)

	Liiketoi-mintajohto	Liiketoi-minta	Sovellus-vastaava	Datan hal-linnan joh-taja	Datan laadun johtaja	Datavas-taava	Data-analyy-tikko	Järjestelmä-kehittäjä	Operatiivinen henkilöstö
Datan laatuvirheiden raportointi		CI	CI	A	CI	R			I
Datan laatuvirheiden seuranta				A	I	R	C	I	I
RCA			CI		A		R	CI	
Datan korjaus		CI	CI	I	A	R	CI	CI	
Prosessin kehittäminen		I	A	I	I	C	C	CI	CI
Datan standardisointi ja korjaaminen		I	C	I	A			C	
Ratkaisun tunnistami-nen		A	CI	CI	R	C	C	CI	I
Datan ehostaminen		A	R	I	C	C	C	CI	