Haaga-Helia
ammattikorkeakoulu Oy

Datanhallinta tekoälyn näkökulmasta – opas organisaation kyvykkyyden arviointiin

Satu Etelälahti

Master opinnäytetyö
Liiketoiminnan teknologiat
2021

Tiivistelmä

Haaga-Helia ammattikorkeakoulu Oy

Tekijä(t)
Satu Etelälahti

Koulutusohjelma
Liiketoiminnan teknologiat

Raportin/Opinnäytetyön nimi
Datanhallinta tekoälyn näkökulmasta – opas organisaation kyvykkyyden arviointiin

Sivu- ja liitesivumäärä
78 + 3

Tämä opinnäytetyö on osa Turun Yliopiston luotsaamaa ja Business Finlandin rahoittamaa AIGA-hanketta – The Artificial Intelligence Governance and Auditing. Vuonna 2021 laaditun opinnäytetyön tarkoituksena on tarkastella datanhallinnan merkitystä tekoälykehityksessä ja luoda Loihde-konsernin datanhallinnan konsulteille asiakastyön tueksi tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli, jolla voidaan kartoittaa organisaatioiden datanhallinnan eri osa-alueiden kyvykkyyttä valjastaa liiketoimintadataa tekoälyn käyttöön. Tutkimus ei ota kantaa datanhallinnan lisäksi muihin tekoälykehityksessä vaadittuihin kyvykkyyksiin, kuten teknologia- ja prosessikyvykkyyksiin.

Teoriaosuudessa tarkastellaan tekoälykehityksen kannalta relevantteja datanhallinnan osa-alueita ja hyvän datanhallinnan tuomia hyötyjä organisaatioille. Tämän lisäksi käydään läpi tekoälykehityksen vaiheet ja niihin liittyviä datanhallinnan aktiviteetteja. Teoriaosuuden lopuksi kartoitetaan erilaisia datanhallinnan maturiteettimalleja, joita voidaan soveltaa tekoälyä hyödyntävän organisaation datanhallinnan maturiteetin arvioinnissa.

Opinnäytetyön tutkimusosuudessa kartoitetaan haastattelujen ja ideointityöpajan kautta datanhallinnan ja tekoälyn asiantuntijoiden näkemyksiä niistä datanhallinnan osa-alueista, joiden kehittäminen tietylle maturiteettitasolle on joko ennakkoedellytys onnistuneelle tekoälykehitykselle tai joita tulisi kehittää tietylle maturiteettitasolle tekoälykehityksen aikana, kun tavoitteena on tuotantokelpoinen ja liiketoimintahyötyä tuova tekoälyratkaisu.

Tutkimuksen tulosten perusteella tekoälykehityksen onnistuminen on vahvasti riippuvainen riittävästä datanhallinnan maturiteetista. Tekoälykehitykseen lähdettäessä tarvitaan ennakoivaa datanhallinnan maturiteettia lähes kaikkien datanhallinnan osa-alueiden osalta, koska tekoälyn toiminta on kytköksissä sen hyödyntämään dataan. Hallitulla datan laadun hallinnalla, datavarastojen ja analytiikan hallinnalla sekä datan hallinnoinnilla varmistetaan säädösten mukainen, tuotantokelpoinen ja liiketoimintahyötyä tuova tekoälyratkaisu. Lisäksi tutkimuksessa selvisi, että datanhallinnan maturiteetin arviointiin voidaan soveltaa olemassa olevia ja hyväksi todettuja maturiteettimalleja tietyin painotuspiste-eroin.

Opinnäytteen tuloksia voidaan hyödyntää arvioinnin tukena, kun organisaatiot haluavat selvittää datanhallinnan kyvykkyyttä tekoälykehitystä ajatellen. Tuloksia voidaan lisäksi hyödyntää sekä asettamaan datanhallinnan tavoitematuriteetti sille tasolle, joka palvelee parhaiten tuotantokelpoisen ja aidosti liiketoimintahyödyllisen tekoälyratkaisun kehittämistä, että määrittämään datanhallinnan kehitysaskeleet, jotka parhaiten tukevat tekoälykehitystä.

Asiasanat
datanhallinta, tekoälykehitys, datan hallinnointi, maturiteettimalli, maturiteettianalyysi

Sisällys

1 Johdanto

”Ilman kunnollista datan hallintaa tekoäly on tekoääliö.” (Ilveskero 2021).

Digitalisaation painopiste on vahvasti datassa ja sen hyödyntämisessä liiketoiminnan tarpeisiin. Data on niin analytiikan kuin nyt myös tekoälyn polttoainetta. Mitä enemmän liiketoiminta asettaa moninaisempia tarpeita datan hyödyntämiselle, sitä enemmän se luo painetta panostaa enemmän myös datanhallinnan eri osa-alueisiin. Puutteet datan hallinnoinnissa voivat kostautua tekoälysovellusten kautta vakavimmillaan tuntuvina sanktioina ja brändihaittana tai vähintään siinä, että huonolaatuisen datan takia tekoälystä ei ole mitään hyötyä. Koko organisaation laajuinen ymmärrys datan merkityksestä ja datanhallinnan käytäntöjen omaksuminen osaksi päivittäistä työtä ja organisaatiokulttuuria voi parhaimmillaan johtaa innovaatioihin ja selkeään kilpailuetuun.

Tämä opinnäytetyö käsittelee datanhallintaa tekoälykehityksen näkökulmasta. Hyvä datanhallinta ja datan hallinnointi ovat keskeisessä roolissa, kun organisaatioissa pohditaan valmiutta tarttua tekoälyn tarjoamiin liiketoimintamahdollisuuksiin tai halutaan skaalata tekoälyn hyödyntämistä laajemmin organisaatiossa. Tekoälyn hyödyntämisessä on omat riskialueensa, joita voidaan hallita paremmin kehittyneemmän datanhallinnan avulla. Organisaatioiden on hyvä pystyä arvioimaan, millä tasolla heidän datanhallintansa on ja mitä datanhallinnan osa-alueita heidän tulisi kehittää sekä ennen tekoälykehitykseen lähtemistä että sen aikana, jotta mahdollistetaan ja varmistetaan datanhallinnan osalta säädösten mukainen ja tuotantokelpoinen tekoälyratkaisu sekä sen tuoma arvo liiketoiminnalle.

Tämän opinnäytetyön tarkoituksena on tarkastella datanhallinnan merkitystä tekoälykehityksessä ja luoda kerätyn aineiston perusteella tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli, jonka avulla organisaatiot voivat kartoittaa datanhallintansa eri osa-alueiden kyvykkyyttä edetä tekoälykehityksessä alkuun ja kohti tuotantokelpoista ja skaalautuvaa tekoälyratkaisua.

Opinnäytetyö on osa Turun Yliopiston luotsaamaa ja Business Finlandin rahoittamaa AIGA-hanketta – The Artificial Intelligence Governance and Auditing, jossa tutkitaan ja kehitetään eri organisaatioiden kesken tekoälyn hallintamalleja ja -mekanismeja, joiden tarkoituksena on vähentää tekoälysovelluksia kehittävien organisaatioiden riskejä. Hankkeen tavoitteena on kaupallistaa hallintamalleja ja mekanismeja sekä viedä niitä kansainvälisille markkinoille, jossa hallintamalleille on tilausta tekoälyn tuottamien päätös-

ten läpinäkyvyyden lisäämiseksi. (Turun Yliopisto 2020). Opinnäytetyön kirjoittajan työnantaja, muun muassa digitaalista palvelumuotoilua ja datanhallinnan konsultaatiota tarjoava Loihde-konsernin jäsen Loihde Advisory Oy on mukana AIGA-hankkeessa yhtenä yritysjäsenenä. Loihde Advisory Oy osallistuu erityisesti AIGA-hankkeen AI governance -osion hallintamallien kehitystyöhön, johon myös tämä opinnäyte liittyy. Tämä uudistamisperusteinen kehittämistyö tukee myös Loihde-konsernin datanhallinnan asiantuntijoita valmistautumaan laajenevan ja kehittyvän tekoälyteknologian maailmaan, jonka kaikkia mahdollisuuksia ei vielä tunneta.

1.1 Tutkimuksen tavoitteet ja rajaukset

Opinnäytetyön tavoitteena on rakentaa organisaation datanhallinnan kyvykkyyttä mittaava tekoälykehityksen mukaan painotettu maturiteettimalli, jolla voidaan tarkastella niin tekoälyhankkeen edellytyksiä datanhallinnan kannalta kuin tekoälyhankkeen aikaistakin datanhallintaa. Kehittämistehtävän tavoite ja tutkimuskysymykset on kuvattu peittomatriisissa (taulukko 1), jossa kysymykset ovat kytketty sekä tietoperustaan, haastattelukysymyksiin (liite 1) että tutkimuksen tuloksiin.

Taulukko 1. Kehittämisprojektin tavoite ja tutkimuskysymykset

Tutkimuskysymykset	**Tietoperusta** (kappalenro)	**Haastattelu-kysymykset** (kysymyksen nro)	**Tutkimuksen tulokset** (kappalenro)
K1. Mitä on hyvä datanhallinta?	2.1 Datanhallinnan osa-alueet ja hyödyt	1, 3, 5	4.1; 5
K2. Millainen rooli datanhallinnalla on tekoälykehityksessä?	2.2 Tekoälykehitys ja datanhallinta	2, 6	4.2; 5
K3. Miten tekoälykehitykseen liittyvän datanhallinnan maturiteettia voidaan arvioida?	2.3 Datanhallinnan maturiteettin arviointi 2.4 Tekoälykohtaisen datanhallinnan maturiteetin arviointi	4 7, 8a, 8b	4.3; 5; 5.1

Opinnäytetyössä keskitytään datanhallinnan kyvykkyyksiin nykymuotoista kapeaa tekoälyä hyödynnettäessä tai harkittaessa sen hyödyntämistä eri organisaatioissa. Opinnäytetyö ei käsittele usein datanhallinnan kanssa keskenään sekoitettua informaatioteknologian hallintaa, joten työstä on rajattu pois jälkimmäiseen sisältyvät vaaditut tekniset kyvykkyydet tekoälykehityksessä. Lisäksi työn ulkopuolelle on rajattu muut tekoälykehityksessä vaaditut, datanhallinnan ulkopuoliset strategia-, prosessi- ja kompetenssikyvykkyydet.

Painotettu datanhallinnan maturiteettimalli rakennettiin kohdennettuna yksityiselle sektorille. Kunnallisen sektorin mahdollisten datanhallinnan erityispiirteiden tarkastelu rajattiin opinnäytetyön kehittämistehtävästä pois. Datanhallinnan maturiteettimallin osalta työssä keskityttiin maturiteettimallin rakenteeseen ja sisältöön, jolloin työn ulkopuolelle rajattiin prosessi maturiteetin analysoimiseksi. Lisäksi projektin aikatauluun liittyvistä syistä tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin kehittämiskierrosten lukumäärä rajoitettiin kahteen, sisältäen haastattelukierroksen ja ideointityöpajan, jossa jälkimmäisessä arvioitutettiin maturiteettimallin aihealueiden alustavia painotuspisteitä ennen mallin viimeistelyä.

1.2 Käsitteet

Käsite	Selite
datanhallinta – data management	Niitä käytänteitä sekä niiden käytänteiden ja ohjeistusten kehittämistä, jalkauttamista ja monitorointia, jotka tähtäävät arvoa tuottavaan dataan läpi sen elinkaaren (Sebastian-Coleman 2018, 19).
datan hallinnointi – data governance	Päätöksentekorakenne datanhallinnan lainsäädännölliselle, oikeudelliselle ja toimeenpanevalle toiminnalle (Sebastian-Coleman 2018, 61).
maturiteettianalyysi	Lähestymistapa, jolla kyvykkyyksiä kehitetään perustuen valittuun maturiteettimalliin (Sebastian-Coleman 2018, 46).
maturiteettimalli	Viitekehys, joka määrittelee valittujen kohteiden hallinnoinnin määrän kasvukehityksen maturiteettiasteikon avulla. Käytetään organisaation kyvykkyyksien suunnitelmallista kehittämistä varten. (Sebastian-Coleman 2018, 42–46).
tekoälykehitys	Vaiheistettu prosessi tekoälyratkaisun toteuttamiseksi (Coveyduc & Anderson 2020).

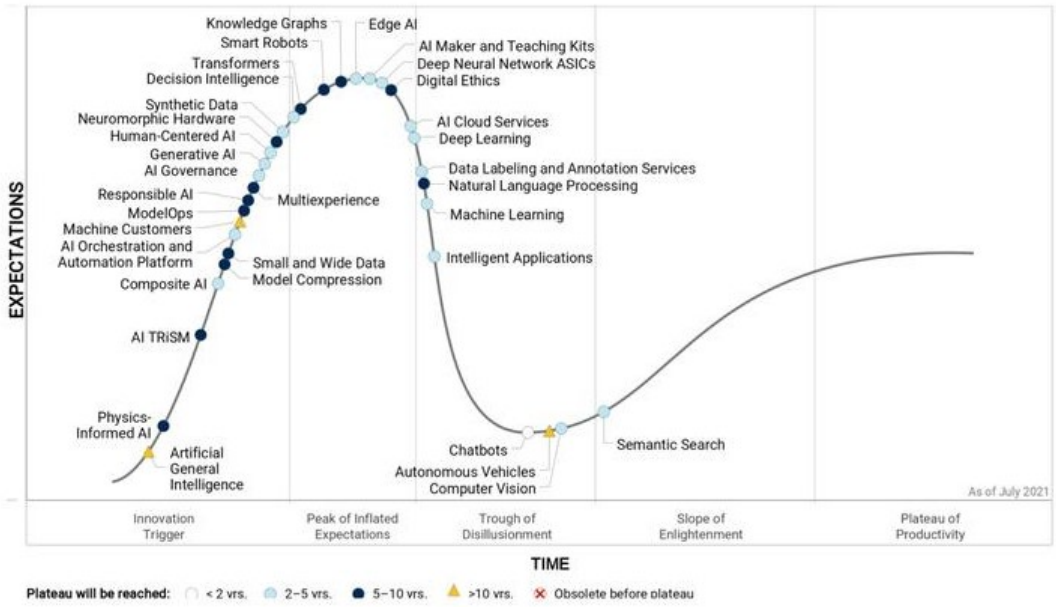
2 Datanhallinnan rooli tekoälykehityksessä

Datan arvo liiketoiminnalle kasvaa nykyistäkin suuremmaksi, kun muun muassa dataa tarvitsevat tekoälysovellukset avaavat uusia liiketoimintamahdollisuuksia. Kuten Gupta ja Mangla (2020) kertovat, tekoälyä hyödynnetään sekä helpottamaan ihmisen työtä, että paikkaamaan ihmisen toimintaa niiltä osin, mihin ihmisen älykkyys ei kykene. Tekoäly ei tarvitse taukoja, se kykenee vaativiin laskutoimituksiin lähes virheittä ja löytää meille nopeimman reitin paikasta A paikkaan B lukemattomien reittimahdollisuuksien joukosta. Tekoäly voi auttaa pelastamaan henkiä identifioimalla vaarallisen yhdistelmän potilaalle määrättyjä lääkkeitä eikä tekoäly ole altis stressin aiheuttamille virheille tai kyvyttömyydelle suorittaa tehtäväänsä vaativissa, monimutkaisissa ja pitkäkestoisissa tilanteissa. Lisäksi tekoäly kykenee muun muassa tunnistamaan kasvoja kuvista, identifioimaan kuvista henkilöitä ja ymmärtämään puhettamme.

Koska tekoälyn toiminta perustuu sen oppimaan ja syötteenä saamaan dataan, AI:n hallinnointiin sisältyy vahvasti datan hallinnointi ja datanhallinta, mihin tässä opinnäytetyössä keskitytään. Tekoälyn perustana olevan datan täytyy olla tarkoitusta vastaavassa kunnossa niin sisällöltään kuin laadultaankin, joten datan hallintaan tarvitaan siten enemmän resursseja. Jos mitä tahansa muuta liiketoimintaomaisuutta, kuten esimerkiksi rahaa ja patentteja hallittaisiin ilman strategiaa, sovittuja toimintatapoja, rooleja ja vastuita sekä ylipäätään ilman ymmärrystä liiketoimintaomaisuuden arvosta ja yleisistä eettisistä liiketoiminnan periaatteista, tällaisen organisaation toimintaa tuskin saataisiin pitkällä aikavälillä kannattavaksi tai edes lainmukaiseksi. Jos oletetaan, että organisaatio ja siinä työskentelevät ihmiset pyrkivät lähtökohtaisesti toimimaan eettisesti ja liiketoiminnan vakiintuneita käyttäytymistapoja noudattaen, he pyrkivät varmasti toimimaan siten myös datan ja sen eri sovelluskohteita, esimerkiksi tekoälyä hyödynnettäessä. Tekoälyä ei siis tule pelätä, demonisoida eikä toisaalta myös kuvitella, että se yksin avaa taivaan auki rikkauksille. Kun opitaan, mitä tekoäly oikeasti on ja että se pohjimmiltaan on matemaattisia sääntöjä, joita yhdistelemällä dataan saadaan tuloksia ja päätelmiä tuottavia tekoälymalleja, voidaan siirtyä ei niin tieteisfiktiolta kuulostaviin termeihin kuten datanhallinta ja datan hallinnointi. Kun näihin liittyvät osa-alueet ovat riittävällä maturiteettitasolla, perusta tuotantokelpoisen tekoälyratkaisun kehittämiselle on olemassa. Terminaattori-uhkakuvat ovat siis tästä vielä kaukana, joten tekoälyn hyödyntämistä voidaan pohtia samanlaisilla tavoilla kuin mitä muuta datalähtöistä liiketoimintaideaa tahansa.

Miksi sitten organisaatiomme pitäisi keskittyä datanhallinnan kehittämiseen juuri nyt, jos haluamme vain kokeilla tekoälyn mahdollisuuksia siellä täällä testihankkeiden kautta? Kuten kuvasta yksi havaitaan, useat tekoälysovellusalueet tulevat Gartnerin analyysin

mukaan leviämään markkinoille ja siirtymään hypekäyrällä ensimmäisestä kehitysvaiheesta käyrän loppuvaiheeseen eli tuottavuuden tasangolle seuraavien 2–5 vuoden aikana. Yksi näistä tekoälyn sovellusalueista on AI:n hallinnointi, tutummin AI governance, jolla varmistetaan, että ihmiset ovat ajan tasalla siitä, miten ja miksi tekoäly toimii, kuten se toimii ja että vastuuroolit ovat paikallaan riskien hallitsemiseksi. (Combs 2021).



Kuva 1. Tekoälyn hypekäyrä, 2021 (Gartner 2021)

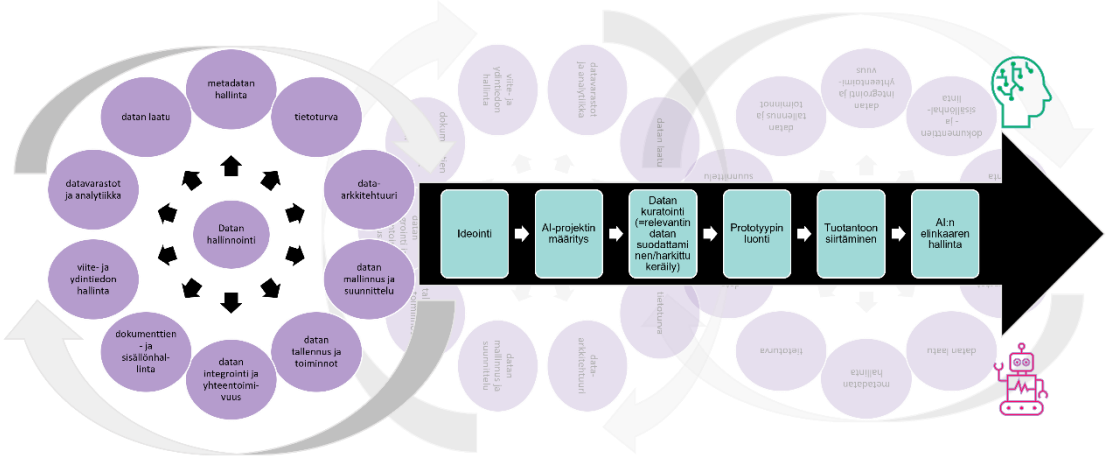
Koska perusteet tekoälyn toiminnalle piiloutuvat muun muassa tekoälyn taustalla olevaan dataan, tekoälyn toimintaa ja sen käyttämässä datassa tapahtuvia muutoksia on valvottava tarkemmin. Useat valtiot ja esimerkiksi Euroopan Unioni ovat vastaamassa tähän tarpeeseen valvoa ja säännellä tekoälyn käyttöä. Euroopan komissio on ilmoittanut (2021), että Euroopan parlamentti ja neuvosto on tehnyt asetusehdotuksen tekoälyn harmonisoidusta sääntelystä – Artificial Intelligence Act, joka on oikeudellinen viitekehys luottamuksenarvoisen tekoälyn kehittämiselle ja hyödyntämiselle. Tämän tavoitteena on se, että EU:n kansalaiset voivat luottaa EU:n markkinoilla olevan ja/tai EU:n kansalaisiin vaikuttavan teknologian turvallisuuteen, lainmukaisuuteen ja perusoikeuksien kunnioittamisen toteutumiseen.

Lisäksi tekoälyn hallinnoinnille löytyy ylätason periaatteisiin keskittyvä, mutta vielä kehityksen alla oleva kansainvälisen standardoimisjärjestön ISO:n – International Organization for Standardization julkaisema ISO 38507 -standardi, joka on tarkoitettu joko tekoälyä jo käyttäville tai tekoälyn hyödyntämistä harkitseville organisaatioille. Standardin sisältö on kohdistettu tekoälyn hallinnointifoorumeille, joita

kehotetaan hyödyntämään myös muita tarkoituksenmukaisia standardeja toimintansa tukena. ISO 38507 -standardi painottaa ihmisen toiminnan ja ihmislähtöisen hallinnoinnin roolia tekoälyn kehittämisessä ja hyödyntämisessä teknologialähtöisyyden sijaan. (ISO/IEC DIS 38507:en 2021).

Kun tekoälyteknologian sovellusmahdollisuudet kasvavat ja useammat organisaatiot kiirehtivät tekoälyratkaisujen kehittämiseen, datanhallinnan kypsyystaso voi olla jopa ratkaiseva tekijä siinä, mitkä organisaatiot jäävät tekoälykehityksessä kokeiluasteelle ja mitkä saavat kehitettyä tuotantokelpoisia, arvoa tuottavia ja säädösten mukaisia tekoälyratkaisuja. Jos tavoittelee joko organisaationlaajuista tai esimerkiksi tietyn liiketoiminta-alueen laajuista AI-valmiutta, on syytä selvittää organisaation datanhallinnan eri osa-alueiden kypsyystaso valitulla tekoälyn vaikutus- ja toiminta-alueella. Maturiteettianalyysissa identifioitujen datanhallinnan eri osa-alueiden kypsyystasojen kautta voidaan määrittää seuraavat tärkeät askeleet kohti AI-valmista organisaatiota datanhallinnan osalta.

Datanhallinta ja sen osa-alueet voidaan ajatella rattaana (kuva 2), joka pyörii läpi tekoälykehityksen, jonka lopputulos on vain niin hyvä kuin tekoälyratkaisuun liittyvän datan hallinta ja hallinnointi eli relevantit rataspyörän osa-alueet ovat. Yhtenä tutkimuskysymyksenä opinnäytetyössä pyritäänkin vastaamaan siihen, mitä hyvä datanhallinta tarkoittaa käytännön tasolla ja miten se heijastuu organisaatioiden toimintaan.



Kuva 2. Datanhallinnan rooli tekoälykehityksessä (mukaillen Sebastian-Coleman 2018, Coveyduc & Anderson 2020. Etelälahti 2021)

Kuvassa kaksi on kuvattu tekoälykehityksen tyypillisimmät vaiheet mustan nuolen päällä ja kansainvälisen tiedonhallinnan järjestön DAMA:n mukaisesti jaotellut datanhallinnan

osa-alueet lilan värisellä rataspyörällä. Nämä kaksi aihealuetta, joita tässä opinnäytetyössä käsitellään, on tuotu kuvassa yhteen visualisoimaan myös tämän työn ydintä eli onnistuneen tekoälykehityksen riippuvuutta hyvästä datanhallinnasta. Toisen tutkimuskysymyksen osalta opinnäytetyön tavoitteena on vastata tarkemmin siihen, millainen rooli datahallinnalla on tekoälykehityksessä.

Jos datanhallinnan rataspyörää ei öljytä eli huolleta oikeista kohdista oikea-aikaisesti, ajan saatossa kasvaa riski sille, että tekoälyn sijaan päädytään niin sanottuun tekoääliöön eli tuotantokelvottomaan tai hyödyttömään tekoälyratkaisuun. Jos taas rataspyörän eri osista pidetään huolta eli datanhallinnan osa-alueita kehitetään oikea-aikaisesti ja tarvittavalle maturiteettitasolle läpi tekoälykehityksen, päädytään varmemmin liiketoiminnalle arvoa tuottavaan tekoälyratkaisuun. Datanhallinnan osa-alueiden maturiteettitasovaatimuksissa voi kuitenkin olla eroja riippuen muun muassa tekoälykehityksen vaiheesta. Maturiteettitason ei kuitenkaan tarvitse olla korkeimmalla mahdollisella tasolla heti ideointivaiheeseen lähdettäessä vaan tekoälykehityksen eri vaiheissa voi olla tarpeen keskittyä eri datanhallinnan osa-alueisiin ja kehittää näiden maturiteettia tarpeen mukaan matkan varrella.

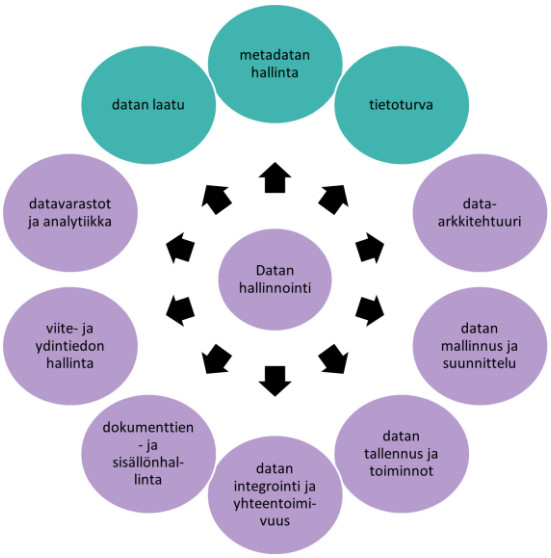
Opinnäytetyön kolmantena tutkimuskysymyksenä pyritään vastaamaan siihen, miten tekoälykehitykseen liittyvän datanhallinnan maturiteettia voidaan arvioida. Jotta pystytään vastaamaan tähän kysymykseen ja rakentamaan tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli, on tarpeen edelleen selvittää, mikä tulisi olla organisaation datanhallinnan tavoitematuriteettitila AI-valmiuteen eli kykyyn siirtää tekoälyratkaisu tuotantoon ja miten eri datanhallinnan osa-alueiden kehitystä tulisi painottaa eli priorisoida tekoälykehityksen ideointivaiheesta AI:n elinkaaren hallintaan saakka.

Kehitettävät datanhallinnan osa-alueet ja niiden tarkemmat sisällöt valikoituvat myös sen mukaan, miten laajasti tekoälyä halutaan hyödyntää ja mistä kaikkialta tekoälyä varten halutaan hakea dataa. Tässä opinnäytetyössä tarkastellaan tekoälykehityksen laajuudesta riippumatta, mitkä datanhallinnan osa-alueet tulisi ottaa huomioon missäkin tekoälykehityksen vaiheessa ja mille maturiteettitasolle kehittää, jotta voidaan puhua AI-valmiista organisaatiosta, jolla on edellytys edetä kehityskaarella eteenpäin. Datanhallinnan osa-alueiden kehitys tulee siis skaalata tekoälykehityksen toiminta- ja vaikutusalueen mukaiseksi.

2.1 Datanhallinnan osa-alueet ja hyödyt

Tässä osiossa käydään läpi datanhallinnan osa-alueet, keskeisimmät standardit ja regulaatiot sekä hyvästä datanhallinnasta koituvat hyödyt organisaatioille datanhallinnan osa-alueittain. Tämän opinnäytetyön tavoitetta ajatellen on tärkeää pohtia, mitkä datanhallinnan osa-alueet ovat tärkeä osa tekoälykehitystä sekä miten datanhallinnan maturiteettitasovaatimukset jakautuvat eri datanhallinnan osa-alueiden välillä, kun halutaan valjastaa tekoäly liiketoiminnan käyttöön.

Organisaatioiden liiketoimintaprosesseissa virtaa dataa enemmän kuin koskaan. Jotta tämä data olisi luotettavaa ja asianmukaisesti saatavilla, tarvitaan datanhallintaa eli suunnittelua, prosesseja, hallinnointia sekä johdon ja koko organisaation sitoutumista asetettuihin tavoitteisiin. Datanhallinta on joukko eritasoisia aktiviteetteja hyvin teknisistä tehtävistä lähtien aina strategisen tason suunnitteluun saakka. Datanhallinnan aktiviteetit ovat niiden käytänteiden ja ohjeistusten kehittämistä, toteuttamista ja seurantaa, joilla hallinnoidaan arvoa tuottavaa dataa läpi sen elinkaaren. Datanhallinta voidaan pilkkoa yhteentoista alueeseen: datan hallinnointi, data-arkkitehtuuri, datan mallinnus ja suunnittelu, datan tallennus ja toiminnot, tietoturva, datan integrointi ja yhteentoimivuus, dokumenttien- ja sisällönhallinta, viite- ja ydintiedon hallinta, datavarastot ja analytiikka, metadatan hallinta ja datan laatu. (Sebastian-Coleman 2018, 1–20). Nämä osa-alueet esitetään usein ympyräkaaviossa, jossa datan hallinnointi on asetettu kaiken keskiöön (kuva 3).



Kuva 3. Datanhallinnan osa-alueet (mukaillen Sebastian-Coleman 2018)

Datanhallinnan osa-alueet voidaan edelleen jaotella datan elinkaaren aikaisiin aktiviteetteihin ja hyvän datanhallinnan perustan rakentaviin aktiviteetteihin, jotka luovat edellytykset johdonmukaiselle datan elinkaaren hallinnalle. Datan elinkaari koostuu kaikista muun muassa datan luomiseen, käyttöön, muokkaamiseen, jakamiseen ja siirtämiseen liittyvistä datan hallinnan prosesseista. Datanhallinnan perustan luovat aktiviteetit tulee huomioida jo osana datan hallinnan ja hallinnointirakenteen suunnitteluvaihetta. (Sebastian-Coleman 2018, 21–33). Kuvassa kolme elinkaaren aikaiset aktiviteetit on esitetty lilalla taustavärillä, ja datanhallinnan perustan luovat aktiviteetit vihreällä taustavärillä.

Hyvien datanhallinnan käytäntöjen varmistama hyvälaatuinen data tuo useita hyötyjä. Laadukkaalla datalla voidaan muun muassa parantaa asiakaskokemusta, nostaa tuottavuutta, mahdollistaa nopean liiketoimintamahdollisuuksiin reagoinnin ja antaa kilpailuetua datasta nousseiden oivallusten kautta. Datan arvon ymmärtämiseksi organisaatiossa voidaan myös laskea huonolaatuisesta datasta johtuvat kustannukset. Hyvä datanhallinta on myös riskienhallintaa. (Sebastian-Coleman 2018, 31–150).

Datan hallinnointi, yleisemmin tunnettu käsitteenä ’data governance’, on kaiken datanhallinnan keskiössä. Datan hallinnointi muodostaa päätöksentekorakenteen, jossa datalle ja datanhallinnan aktiviteeteille osoitetaan tarvittavat roolit vastuineen ja päätöksentekooikeuksineen. Hallinnointi on sekä lainsäädännöllistä, oikeudellisia että toimeenpanevaa toimintaa. Datan hallinnointiin liittyy niin datan arvon määrittäminen kyseiselle organisaatiolle, data strategian luominen kuin datan hallinnan käytäntöjen asettaminen sekä niiden toteutumisen seuraaminen ja maturiteetin kehittäminen. Datan hallinnointi on siis jatkuvaa, organisaation prosesseihin sulautettua toimintaa ilman päätepistettä. Tähän liittyy tiukasti tarve organisaation kulttuurin muutokselle kohti parempaa ymmärrystä datan hallinnasta ja sen tuomasta arvosta. (Sebastian-Coleman 2018, 20–67).

Datan hallinnoinnille löytyy kansainvälisen standardoimisjärjestön julkaisema ISO 38505 -standardi, jonka tarkoitus on tukea datan hallinnointifoorumin ja organisaation ylimmän johtoryhmän välistä kommunikointia, jotta varmistetaan, että datanhallinta on linjassa organisaation strategian kanssa. ISO 38505 -standardi käsittelee sellaisen tiedon identifiointia, jota hallinnointifoorumi tarvitsee arvioidakseen ja ohjatakseen datalähtöisen liiketoiminnan käytäntöjä ja suuntaa. Lisäksi standardi auttaa identifioimaan niitä kyvykkyyksiä ja työkaluja, joita tarvitaan datan ja sen käytön monitorointia varten. (ISO/IEC TR 38505:en 2018).

Datan hallinnointi on läpi organisaation eri liiketoimintojen virtaavalle datalle ja sen käytölle asetettuja yhteisiä viitekehyksiä ja ohjeistuksia, jotta data olisi yhdenmukaista ja josta

siten voidaan tehdä kokonaisvaltaisia ja johdonmukaisia organisaatiotason päätöksiä. Datan hallinnoinnin hyöty tuleekin jo siitä, että se opastaa toimintaa kaikilla muilla datanhallinnan osa-alueilla. Yhdenmukaisen datanhallinnan kautta dataan liittyviä päätöksiä voidaan tehdä linjassa liiketoimintastrategian kanssa sen sijaan, että päätöksiä tehtäisiin projektikohtaisesti. Keskitetysti laaditut ja kommunikoidut datan laatu- ja käyttösäännöt antavat työntekijöille selkeät ohjeet datan käytölle ilman tarvetta tehdä määritelmiä aina uudestaan ja lisäten luottamusta siihen, että data on hyvänlaatuista. Lisäksi mitä enemmän säätelyä kohdistuu organisaation liiketoimintaan, sitä suurempaa hyötyä nähdään datan hallinnoinnilla sekä riskien vähentämisen muodossa kuin prosessien tehostamisenkin kautta. (Sebastian-Coleman 2018, 61–78).

Datan elinkaaren aikaisia aktiviteetteja tukevat sekä hyvä tietoturva, metadatan hallinta, että riittävä datan laatu. Näitä kaikkia kolmea datanhallinnan osa-aluetta tulee kehittää läpi datan elinkaaren, jotta organisaatiossa varmistetaan datan luotettavuus ja sitä kautta datasta saatava arvonnousu. Tietoturvan, metadatan ja datan laadun hallinta ovat datan hallinnoinnin peruspilareita, jotka tulee integroida osaksi organisaation prosesseja. (Sebastian-Coleman 2018, 14–24).

Metadatan hallinta keskittyy datasta informaatiota antavan datan, kuten määritelmien ja lähdejärjestelmätiedon hallintaan. Hyvin käytäntöjen kautta datasta kerätään tietoa, jolla kasvatetaan koko organisaation tietotasoa ja jonka avulla voidaan muun muassa identifioida tarpeetonta dataa ja estää huonolaatuisen tai vanhentuneen datan hyödyntämisen. Metadatan hallinta on ennakkoedellytys onnistuneelle datanhallinnalle ja sen merkitys kasvaa sitä suuremmaksi mitä enemmän organisaatio kerää ja varastoi dataa. Hyvällä metadatanhallinnalla voidaan esimerkiksi varmistaa, että organisaatiossa kyetään identifioimaan henkilökohtaisia ja arkaluonteisia tietoja läpi järjestelmien. Ilman metadataa riskeerataan kyvykkyys hallinnoida organisaation dataa ylipäätään. (Sebastian-Coleman 2018, 20–150). Koska datan hallinnointi tähtää kokonaisvaltaiseen datanhallinnan kehittämiseen, metadatan hallinta on kriittinen työkalu hallinnoinnin ja dataan liittyvien kehittämistavoitteiden onnistumisessa.

Hyvällä tietoturvalla varmistetaan sekä tietosuoja ja tiedon luottamuksellisuuden pysyvyys, että oikeanmukaiset pääsyoikeudet tiedolle. Ensiksi identifioidaan suojausta vaativa data ja haarukoidaan järjestelmät, joissa kyseistä dataa on. Tämän jälkeen määritetään suojauksen taso ja identifioidaan ne liiketoimintaprosessit, jotka tarvitsevat kyseistä dataa. Tämän perusteella määritetään, millä perusteilla ja ehdoilla dataa voidaan hyödyntää. (Sebastian-Coleman 2018, 20–139). Mitä paremmin tietoturvariskejä hallitaan, sitä turvallisempaa on laajentaa tekoälyn hyödyntämistä yhdeltä liiketoiminta-alueelta tai -prosessista

laajemmalle. Datamäärien kasvaessa ja hyödynnettäessä dataa aiempaa moninaisemmin tavoin kasvavat paineet datan käytön turvaamiseksi ja säätelemiseksi. Tunnetuin näistä säätelykeinoista on vuonna 2018 voimaanastunut EU:n yleinen tietosuoja-asetus, tutummin GDPR – General Data Protection Regulation.

GDPR säätelee henkilötietojen käsittelyä. Asetus antaa erilaisia asetuksia riippuen missä roolissa henkilötietojen käsittelijäorganisaatio toimii suhteessa henkilötietoon. Näitä rooleja ovat tiedon käsittelijä ja rekisterinpitäjä eli tiedon omistaja. Tiedon omistajan vastuulla on suojella rekisteröimäänsä henkilötietoa tarpeellisilla teknisillä ja organisatorisilla toimilla, esimerkiksi asettamalla datalle kontrollipisteitä, jolloin organisaatio toimii jo suunnitteluvaiheessa oletusarvoisesti tietosuoja-asetuksen edellyttämällä tavalla. Tiedon käsittelijän vastuulla on taasen käsitellä tiedon omistajan omistamaa henkilötietoa sen vaatimalla tavalla. (IT Governance Privacy Team 2020). Kun puhutaan eettisestä datanhallinnasta, regulaatiot toki tukevat eettisyyden toteutumista organisaatioissa, mutta mitä kehittyneempää analytiikkaa ja teknologiaa datanhallinnassa käytetään, sitä enemmän tarvitaan kuitenkin myös organisaatiokulttuurin muutosta ulkopuolelta tulevan kontrollin lisäksi. Tällöin eettinen tapa toimia on osa normaalia, vakiintunutta tapaa toimia eikä vain pakotettu paha, johon kiinnitetään huomiota vasta, kun ollaan vaarassa jäädä rikkeestä kiinni. Kokonaisvaltaisesti luotettavat eli sekä lainmukaisesti, eettisesti että kestävästi tekoälyä hyödyntävät organisaatiot voittavat myös enemmän asiakkaita puolelleen. (Sebastian-Coleman 2018, 49–60). Kehittynyt datan hallinnointi ja tämän tuoma ajattelun ja lopulta organisaatiokulttuurin muutos on se elementti, jonka kokonaisuudessaan voi ajatella varmistavan sekä yksilön, tiimien että koko organisaation toiminnan eettisyyden muun muassa tekoälyn kehityksen ja hyödyntämisen suhteen.

Kun organisaatio haluaa datasta liiketoiminnalleen arvoa, datan on oltava datan hyödyntäjien näkökulmasta laadukasta. Datan laadukkuutta voidaan arvioida erilaisten ulottuvuuksien kautta. Datan ulottuvuudet mittaavat yleensä datan täydellisyyttä eli onko dataa ylipäätään tarpeeksi, datan oikeellisuutta eli datan tarkkuutta ja validiutta, datan yhteentoimivuutta eli kuinka johdonmukaista, eheää ja ainutlaatuista data on sekä datan ajantasaisuutta, saatavuutta, käytettävyyttä ja tietoturvallisuutta. Datan laatu varmistetaan suunnittelemalla ja jalkauttamalla tekniikoita datan laadun mittaamiseksi, arvioimiseksi ja kehittämiseksi. (Sebastian-Coleman 2018, 20–162). Datan laadun hallinnalle löytyy kansainvälisen standardoimisjärjestön julkaisema ISO 8000 -standardi, jossa informaatiolle ja datalle on määritelty niiden laatua määrittävät ominaisuudet. Lisäksi standardi sisältää menetelmiä informaation ja datan laadun hallintaan, mittaamiseen ja kehittämiseen. (ISO/TS 8000-60:en 2017). Jos data on oikeanmukaista, kattavaa ja

ajantasaista, se on jo siten vähemmän riskialtista ja ylipäätään paremmin hyödynnettävissä. (Sebastian-Coleman 2018, 20–29). On sanomattakin selvää, että datasta oppiva tekoäly antaa sitä relevantimpia tuloksia, mitä laadukkaampaa dataa sille syötetään. Oikeanmukainen data johtaa oikeanmukaisempaan tekoälyyn, ajantasainen data johtaa ajankohtaiseen tekoälyyn ja kattavampi data opettaa tekoälyä paremmaksi toiminnassaan. Esimerkiksi, jos dataa niin sanotusti siivotaan liikaa jättämällä vaikkapa vapaatekstikentät tekoälyä opettavasta datasta pois, data ei välttämättä ole enää tarpeeksi kattavaa ja antaa siksi erilaisia tuloksia kuin jos myös vapaatekstikentät olisi otettu huomioon. Datan laadukkuus pitää kuitenkin määritellä jokaista liiketoimintatapausta kohdin erikseen.

Seuraavaksi läpikäydyt loput datanhallinnan osa-alueista sisältävät datan elinkaaren aikaisia aktiviteetteja, joiden hallinnassa keskitytään organisaation liiketoimintakriittisimpään dataan ja tarpeettoman datan minimointiin. Datan elinkaaren aikaisia aktiviteetteja ovat data-arkkitehtuuri, datan mallinnus ja suunnittelu, datan tallennus ja toiminnot, datan integrointi ja yhteentoimivuus, datavarastot ja analytiikka, viite- ja ydintiedot sekä dokumenttien- ja sisällönhallinta. (Sebastian-Coleman 2018, 21–35).

Data-arkkitehtuuri määrittää sen, miten organisaation tietovarantoja hallitaan ja niiden rakenteita suunnitellaan linjassa organisaation strategian ja sen asettamien tavoitteiden kanssa. Tätä suunnittelua tehdään dataa mallintaen, mikä on eri liiketoiminnan osa-alueiden asettamien datavaatimusten identifiointia, analysointia ja kommunikointia varten käytetty prosessi. Data-artefakteilla eli tietomalleilla, määritelmillä ja tietovirtakuvauksilla saadaan organisaation valtavat datamassat sellaiselle abstraktiotasolle, jota liiketoiminnan johto voi ymmärtää ja jonka perusteella se voi tehdä päätöksiä (Sebastian-Coleman 2018, 23–83). Tarve tällaiselle dokumentaatiolle kasvaa, kun pidetään kirjaa siitä, mitä dataa annetaan tekoälylle.

Data-arkkitehtuurityössä luodaan ja ylläpidetään organisaatiotason tietoa datasta eli metadataa, jonka avulla dataa voidaan hallita arvoa tuovana työkaluna. Data-arkkitehdit tekevät esimerkiksi dataan liittyvää mallinnus- ja suunnittelutyötä, jotta data olisi parhaalla mahdollisella tavalla liiketoiminnan käytettävissä. Suunnittelutyötä tehdään muun muassa dataa mallintaen. Tietomallinnuksen avulla kerättyä arkkitehtuurista dokumentaatiota voidaan hyödyntää, kun esimerkiksi etsitään uusia datan käyttömahdollisuuksia tai halutaan hallita paremmin monimutkaisista ja joustamattomista datarakenteista koituvia riskejä ja kuluja. (Sebastian-Coleman 2018, 85–101). Yritysarkkitehtuurille, joka sisältää myös data-arkkitehtuurin, on olemassa viitekehyksiä, joista yhtenä ensimmäisistä kehitettiin Zachmanin viitekehys vuonna 1987 (Sebastian-Coleman 2018, 82). Muita vastaavia viitekehyksiä ovat muun muassa TOGAF – The Open Group Architecture Framework ja

suomalainen JHS 179 – Julkisen hallinnon suositus 179. Kuten Sebastian-Coleman (2018, 84) kuvaa, organisaation data-arkkitehtuuri kuvataan eri abstraktiotason dokumentaatiolla. Data-arkkitehtuurikuvausten kautta datalle kerätään sille asetettuja vaatimuksia, ohjataan dataintegraatioita ja varmistetaan, että arkkitehtuuri tukee organisaation datastrategiaa. Data-arkkitehtuurilla saavutettava maksimaalinen hyöty on sitä suurempi, mitä laajemmalle se ulotetaan, koska arkkitehtuurilla mahdollistetaan liiketoimintakriittisen datan standardointi ja integrointi läpi organisaation.

Datan tallennus ja toiminnot -osa-alue mielletään perinteiseksi datanhallinnaksi, johon kuuluu tallennettavan tiedon järjestelmäkohtaiset suunnittelu-, jalkautus- ja tukivaiheet datan syntymisestä sen hävittämiseen saakka. Tämä toiminta tukee siis koko datan elinkaarta, tavoitteena datan arvon maksimointi. Näistä teknisistä toiminnoista vastaavat yleensä tietokantojen ja verkkojen ylläpitäjät, joilla on rooli datan hallinnointi -rakenteessa tahoina, joiden kriittistä tietämystä teknisistä ympäristöistä hyödynnetään data- ja liiketoimintavastuullisten toimintaohjeita jalkauttaessa ja toisaalta he voivat myös auttaa uusien teknologioiden omaksumisessa ja hyödyntämisessä. (Sebastian-Coleman 2018, 23–102).

Jos datan tallennus ja toiminnot -osa-alueella keskitytään järjestelmä- ja ympäristökohtaisiin datan ylläpitoon liittyviin aktiviteetteihin, niin datan integrointi ja yhteentoimivuus -puolella aktiviteetit taas liittyvät prosesseihin tiedon siirtämiseksi ja yhdistämiseksi eri tietovarastojen, järjestelmien ja organisaatioiden sisällä ja välillä. Tavoitteena on saattaa tarvittu data saataville oikeassa muodossa ja oikeaan aikaan sekä identifioida tapahtumatiedoista mahdollisuuksia ja uhkia. Kustannushyötyjä sekä prosessitehokkuutta voidaan saavuttaa keskittämällä dataa. Kaikki tämä palvelee esimerkiksi analytiikkaa, jossa halutaan varmistua siitä, että dataa päivitetään oikein, data on sen keräämisen jälkeen nopeasti hyödynnettävissä, data liikkuu ongelmitta eri tietovarastojen välillä ja että prosessit tukevat datan johdonmukaisuutta ja jatkuvuutta. Datan integrointi ja yhteentoimivuus -puoli onkin riippuvainen monista muista datanhallinnan osa-alueista, jotta se voi omalta osaltaan saavuttaa sille osoitetut tavoitteet. (Sebastian-Coleman 2018, 23–105).

Datavarastot ja analytiikka -aktiviteetit tukevat päätöksentekoa varten tarvittavan datan hallintaa keskittyen datan analysoinnin ja raportoinnin kyvykkyyksien kehittämiseen. Datavarastoja tuleekin kehittää sidottuna vahvasti organisaation asettamiin prioriteetteihin, jotta ratkaisut palvelevat liiketoimintaa. Datavarasto rakentuu useasta osasta, joiden läpi data liikkuu ja jonka tuloksena datarakenteet ja datan muoto voivat muuttua riippuen datan käyttötarkoituksesta, olkoon se esimerkiksi raportti tai syöte sovellukseen, jossa dataa analysoidaan. Datavarasto mahdollistaa keskitetyn paikan eri järjestelmistä saadun datan tietoturvalliseen jakamiseen ja analysointiin. Kehittyneemmät datajärvi-ratkaisut

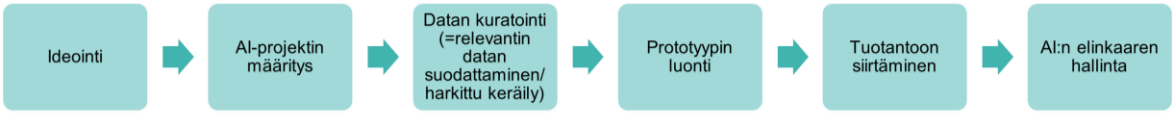
mahdollistavat lisäksi ennustavat analyysit suurten datamäärien tallennusmahdollisuuden ja nopeutensa ansiosta. (Sebastian-Coleman 2018, 23–110).

Viite- ja ydintiedon hallinta on organisaation liiketoiminnalle kriittisen, yleensä laajasti useissa järjestelmissä hyödynnettävän tiedon paikkansapitävyyden, ajantasaisuuden ja merkityksellisyyden varmistamista. Hyvällä ydintiedon, kuten asiakas- ja tuotetiedon hallinnalla sekä hyvällä muuta dataa kategorisoivan referenssidatan, kuten postinumeroiden ja maakoodien hallinnalla kasvatetaan kriittisen datan kattavuutta, oikeellisuutta, ajantasaisuutta, ymmärrettävyyttä ja siten näiden kautta datan luotettavuutta ja hyödynnettävyyttä. Tällaisella datalla saadaan luotettavasti ymmärrystä muun muassa asiakkaista ja tuotteista sekä voidaan tehdä ennustavaa analyysia tulevaisuutta ajatellen. Hyvällä ydintiedonhallinnalla voidaan siis lisätä tehokkuutta ja vähentää riskejä, mitä eroavaisuudet eri järjestelmien datarakenteiden välillä voivat aiheuttaa. (Sebastian-Coleman 2018, 23–114).

Dokumenttien- ja sisällönhallinta sisältää strukturoimattoman datan elinkaaren aikaisen hallinnan. Strukturoimatonta dataa ei voida tallentaa perinteisiin tietokantatauluihin, mutta tällaiseenkin dataan kohdistuu muun datan tapaan vaatimuksia pääsynhallinnan, käytön ja säilytysajan suhteen. Dokumenttienhallinnalla tarkoitetaan sekä sähköisten että paperidokumenttien organisointia ja hallintaa läpi niiden elinkaaren. Sisällönhallintaan kuuluu taas muun muassa dokumenttien, videoiden ja kuvien sisällön kategorisointia, organisointia ja järjestelyä siten, että ne ovat monilla eri tavoin hyödynnettävissä. (Sebastian-Coleman 2018, 23–115).

2.2 Tekoälykehitys ja datanhallinta

Kuten mikä tahansa muukin projekti, myös tie tekoälyn hyödyntämiseen alkaa ideoinnista ja projektin määrittelyvaiheesta (kuva 4). Tätä seuraavat yleensä datan kuratointi, prototyypin luonti, tuotantovaihe sekä lopulta tekoälyn elinkaaren hallinta. Tämä vaiheistus on yleistetty versio tekoälykehityksestä. Organisaatiokohtaisesti vaiheissa voi olla eroja ja tekoälykehitys voidaan myös jakaa iteraatioihin. (Coveyduc & Anderson 2020).



Kuva 4. Tekoälykehitys (mukaillen Coveyduc & Anderson 2020)

Organisaatioiden tulee ympärillä vellovasta tekoälyhuumasta huolimatta tehdä teknologiavalinnat liiketoimintastrategian ja sen datalle antamien vaatimusten pohjalta, ei päinvastoin (Sebastian-Coleman 2018, 37). Siksi jokaisen tekoälyn hyödyntämistä harkitsevan organisaation tulee kysyä itseltään, mitä oikeaa liiketoimintaongelmaa tekoälyllä halutaan ratkaista, miten organisaatio toimii operatiivisesti tällä hetkellä ongelman suhteen ja miten organisaation on mahdollista hyötyä AI-teknologiasta tulevaisuudessa (Coveyduc & Anderson 2020). Olkoon organisaatio ratkaisemassa tekoälyllä jotain liiketoimintaongelmaa tai kartoittamassa uusia mahdollisuuksia, jokainen tekoälyn käyttötapaus tarvitsee toimiakseen dataa. Siispä jo tekoälykehityksen alussa on kriittistä ymmärtää, mitä dataa organisaatiolla on eli datasta tarvitaan sitä kuvaavaa metatietoa ideoinnin tueksi.

A-projektin ideointi on hedelmällisintä, jos organisaatiokulttuuri tukee sitä laajalti. Ideoita voidaan kerätä ideapankkiin esimerkiksi erilaisista analyyseista ja haastatteluista, mutta ideoiden suodattamiseksi tarvitaan hyvin määritelty kriteeristö. Lisäksi ideoiden katselmoinnin ja jalkautuksen on oltava säännöllisesti toistuva prosessi. Organisaatiossa voi olla ideointipankin lisäksi innovointiin keskittynyt työryhmä, jolla on päätäntävaltaa tehdä muutoksia. Päätöksentekoa varten työryhmällä on oltava vaadittua ymmärrystä olemassa olevasta tekoälyteknologiasta ja niiden kyvykkyyksistä. (Coveyduc & Anderson 2020). Tekoälyteknologian ymmärtäminen sisältää myös tekoälyn tarvitseman datan tärkeyden ymmärtämisen. Mitä datalähtöisempi kulttuuri organisaatiossa on, sitä rikkaampia ja realistisempia tekoälyllä toteutettavia ideoitakin organisaatio voi synnyttää.

Kun tekoälyhanke on jonkin idean osalta päätetty toteuttaa, hankkeen alussa tekoälyratkaisulle määritetään tavoitteet ja ne myös priorisoidaan (Thomas 2019). Hankkeelle laaditaan tarkempi projektisuunnitelma ja varmistetaan suunnitelman realistisuus. Lisäksi suunnitelma pilkotaan pienemmiksi mitattaviksi kokonaisuuksiksi, joiden toteutumisen kautta voidaan helpommin seurata tavoitteiden saavuttamista. AI-projektin määritysvaiheessa myös identifioidaan kaikki tekoälyratkaisun osalta tarvittavat sidosryhmät. (Coveyduc & Anderson 2020). Viimeistään tässä vaiheessa hankkeeseen tarvitaan mukaan joukko data-asiantuntijoita, jotka tuovat mukanaan näkymän dataan ja datanhallinnan kyvykkyyteen tekoälyn suhteen.

Mitä kehittyneempää teknologiaa käytämme datan käsittelyyn, erityisesti tekoälyn polttoaineena, sitä tärkeämmäksi tulee ihmisen merkitys sen hallinnoinnissa. Ihmisiä tarvitaan seuraamaan ja selittämään, mitä dataa hyödynnetään ja miksi sekä miten on päädytty mihinkin päätökseen. (Ahopelto 2019). Tarvitaankin nykyistä parempaa yhteistyötä erilaisten data-aktiviteettien parissa työskentelevien ihmisten kesken. Kattavaa, koko

datan elinkaaren aikaista hallintaa varten tarvitaan sekä liiketoimintaymmärrystä, data-arkkitehtuuriosaamista, erittäin vahvaa teknistä osaamista, kykyä analysoida dataa ja siitä tehtyjä löydöksiä, datan mallinnus- ja määritysosaamista yhteisen kommunikointikielen luomiseksi sekä strategista ajattelukykyä uusien datan liiketoimintaa tukevien käyttökohteiden identifioimiseksi. Datan hallinnoinnilla varmistetaan, että kaikki yhteistyön osat eli ihmiset ja prosessit toimivat datan suhteen organisaation tavoitteiden mukaisesti. (Sebastian-Coleman 2018, 38–63).

Tekoälyprojektin määritysvaiheen jälkeen päästään tarkemmin pohtimaan sitä, minkälaista dataa täytyy kerätä käyttötapausten toteuttamista varten. (Thomas 2019). Jo aiemmassa tekoälykehityksen vaiheessa määritetystä tarvittavasta datasta on oltava saatavilla nyt lisäksi metatietoa muun muassa siitä, missä dataa jo tuotetaan tai missä sitä on saatavilla, miten sitä jo käytetään ja miten sen käyttö on suojattu sekä minkä laatuista data on (Sebastian-Coleman 2018, 34). On tärkeää olla tietoinen kaikista saatavilla olevista, niin sisäisistä kuin ulkoisistakin tietolähteistäkin, koska tekoälyhankkeen onnistuminen riippuu osaltaan tekoälyn hyödyntämästä datasta ja sen tarkoituksenmukaisuudesta ja laadusta. (Coveyduc & Anderson 2020). Dataa on erilaista ja sitä voidaan luokitella eri tavoin, kuten esimerkiksi transaktio- ja ydintietoon, eri liiketoiminta-alueen tietoihin tai luottamuksellisuuden mukaan erilaisiin tietoihin. Datan luokasta riippuen siihen kohdistuu erilaisia sen elinkaaren aikaisia vaatimuksia. Tekoälysovellutukset tuovat näihin vaatimuksiin oman lisänsä. Kun organisaatio on määritellyt tietyn datan liiketoimintakriittiseksi tekoälyn hyödyntämisen osalta, tämän tarvittavan datan hallinnan perustan on oltava kunnossa. Kuten minkä tahansa liiketoiminnalle tärkeän datan hallinnassa, myös tekoälyn käyttämän datan hallinnassa tulee ottaa huomioon koko datan elinkaaren aikainen hallinta, datan erityispiirteet ja dataan liittyvän riskien hallinta. (Sebastian-Coleman 2018, 34–44).

Kun tietojoukot on valittu, mietitään, mitä työkaluja ja tekniikoita tarvitaan datan käsittelyä ja varsinaista tekoälymallin rakentamista varten (Thomas 2019). Dataa kerätessä on otettava huomioon muun muassa tietoturvaan ja tietosuojaan liittyvät säädökset, jotta dataa käsitellään lainmukaisesti ja liiketoimintaturvallisesti. Datan hallinnointifoorumi tulee perustaa valvomaan sekä tätä että muuta tekoälykehitykseen liittyvää datanhallinnan toimintaa, jotta varmistetaan näiltä osin, että organisaatio saavuttaa tavoitteensa. Tarvittaville datanhallinnan aktiviteeteille tulee lisäksi laatia operatiivinen suunnitelma. (Coveyduc & Anderson 2020). Jos data ei ole valmiiksi helposti saatavilla, datatieteilijät yhdistelevät tietojoukkoja ja parantavat datan laatua tekoälysovelluksen opettamista varten hyödyntäen erilaisia työkaluja ja alustoja. (Coveyduc & Anderson 2020). Kuitenkin

mitä parempaa datan laatu datan kuratointivaiheessa jo on, sitä nopeammin tekoälyn ammattilaiset pääsevät arvoa tuottavan tekoälyratkaisun kehittämiseen.

Datan kuratoinnin jälkeen valitaan jalkautettavat toiminnallisuudet prototyyppiä varten ja testataan iteratiivisesti, saadaanko tekoälystä odotettua arvoa. Tämän sijaan voidaan myös resurssien säästämiseksi päätyä hankkimaan valmis tekoälyratkaisu, jos markkinoilta löytyy vaatimustenmukainen ratkaisu. Lähestymistapaan vaikuttaa muun muassa se, kuinka kattavaa tietotaitoa organisaatiosta jo löytyy vai tarvitaanko organisaation ulkopuolelta asiantuntijoita avuksi tekoälyratkaisun kehittämistyöhön. Varsinaista tekoälykehitystyötä tehdään yleisimmin iteratiivisesti, hyödyntäen ketteriä menetelmiä ja keräten säännöllisesti palautetta sidosryhmiltä. (Coveyduc & Anderson 2020). Tämän opinnäytetyön tuotoksena kehitetty tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli on kohdistettu organisaatioille, jotka kehittävät tekoälyratkaisuja itse eivätkä pelkästään osta sellaisia markkinoilta valmiina.

AI-prototyypin kehittämisen jälleen pohditaan, miten tekoälymallit siirretään tuotantoon (Thomas 2019). Kun tekoälyratkaisulle rakennetaan jatkuva kehitysputki, on syytä arvioida ensin se, ovatko liiketoiminnan prioriteetit muuttuneet. Tämän jälkeen suoritetaan tekninen arviointi, johon sisältyy muun muassa arvio teknologian kyvystä skaalautua isompiin käyttäjämääriin tai suurempaan datamäärään. Lisäksi rakennetaan tarvittavat käyttäjäsuojausmekanismit ja luodaan testauskehys. Jos ratkaisu on pilvipohjainen, tarvitaan kattavat sopimukset organisaation ja toimittajan välille. Tekoälyratkaisun toiminta tulee rakentaa siten, että automaattisen laaduntestauksen lisäksi ihminen voi puuttua tekoälyn toimintaan ratkaisevissa kohdin. (Coveyduc & Anderson 2020).

Kun tekoälyratkaisu on viety tuotantoon, vuorossa on tekoälyn jatkuvaa monitorointia, jotta varmistutaan tekoälymallin tarkoituksen- ja oikeudenmukaisesta toiminnasta (Thomas 2019). Tekoälyn elinkaaren hallinta on jatkuvaa ja se sisältää myös läpinäkyvää datanhallintaa ja säännöllisiä auditointeja. AI:n elinkaaren hallinnan ohella etsitään keinoja tekoälyn edelleen kehittämiseen, laajentamiseen ja hyödyntämiseen myös muualla organisaatiossa. Tämän sujuvoittamiseksi tekoälyhanke tulee olla riittävän kattavasti dokumentoitu ja viestitty, jotta tekoälymallien jatkohyödyntäminen organisaatiossa on sujuvaa. Lisäksi tarvitaan vastuita datan kehittämiselle, jotta kehitysputkeen saadaan saataville uusia hyvälaatuisia datalähteitä. (Coveyduc & Anderson 2020).

2.3 Datanhallinnan maturiteetin arviointi

Maturiteettiarvioinneilla voidaan mitata organisaation kykyä kehittää toimintaansa tietyllä alueella (Taylor 2020). Datanhallinnan maturiteettimallia käytetään arvioidessa organisaation yhden liiketoiminta-alueen, substanssialueen, yksittäisen prosessin tai idean, tai koko organisaation laajuista datanhallinnan kypsyystasoa. Maturiteettianalyyseja suoritetaan tavoitteena sekä kasvattaa organisaation tietämystä datanhallintansa nykytilan maturiteetista että asettaa seuraavia askeleita kohti organisaation kehittyneempää datanhallintaa. (Sebastian-Coleman 2018, 46–47). Maturiteettianalyysin tuloksia voidaan käyttää arvioimaan datanhallinnan kykyä tukea organisaation strategisia liiketoimintatavoitteita, kehittämään vaadittuja datanhallinnan kyvykkyyksiä, mittaamaan datanhallinnan kehitystä, vertailussa kilpailijoita ja kumppaneita vasten ja kasvattamaan organisaation tietoisuutta datanhallinnan merkityksestä. Askeleet kohti optimaalisempaa maturiteettitasoa vähentävät asteittain huonosta datanhallinnasta aiheutuvia riskejä ja turhaa työtä sekä kasvattavat datan laatua ja tuottavuutta. (Sebastian-Coleman 2020).

Datanhallinnan maturiteettimalleihin sisältyy yleisesti ottaen viisi tai kuusi maturiteettitasoaskelmaa riippuen siitä, onko niin sanottu nollataso asetettu ensimmäiselle askeleelle vai ei (Sebastian-Coleman 2018, 46). Datanhallinnan maturiteettimallit eroavat toisistaan osa-alueiden ja niihin sisältyvän kriteeristön osalta (Sebastian-Coleman 2020). Datanhallinnan maturiteetin arviointiin on kehitetty malleja muun muassa Gartnerilla ja IBM:llä (Taylor 2020). DAMA ei ole kehittänyt valmista datanhallinnan maturiteettimallia vaan antaa ylätason esimerkkikuvauksen eri maturiteettitasoille, jotka pohjautuvat Carnegie Mellon -yliopiston kehittämään CMM – The Capability Maturity Model -malliin. Omaehtoisen maturiteettimallin rakentamiseksi voidaan kuitenkin hyödyntää DAMA-ympyrän tietoaluekohtaista kriteeristöä. Jokaista maturiteettitasoa kohden on joukko kriteerejä tason saavuttamiseksi ja jokainen saavutettu askel näkyy organisaatiossa kyseessä olevan datanhallinnan prosessin johdonmukaisuuden, luotettavuuden ja ennustettavuuden paranemisena. (Sebastian-Coleman 2020).

Taulukossa kaksi (19) on esitetty DAMA:n, Gartnerin ja IBM:n maturiteettimalliasteikkojen ylätason kuvaukset vertailun vuoksi. Tasokohtaisesti esitetyistä maturiteettimallien asteikkojen kuvailuista voidaan havaita paljon yhtäläisyyksiä. Nollatason maturiteetilla tarkoitetaan yleensä tietämättömyyttä datan arvosta ja datanhallinnan periaatteista kaikilla organisaation tasoilla. Voidaan olettaa, että tällä maturiteettitasolla oleva organisaatio ei todennäköisesti ole myöskään havahtunut datanhallinnan maturiteettianalyysin tarpeellisuuteen. IBM ei ole sisällyttänyt nollatasoa omaan datanhallinnan maturiteettiasteikkoonsa.

Taulukko 2. Datanhallinnan maturiteettimalleja (mukaillen Sebastian-Coleman 2018 ja Taylor 2020)

	DAMA	**Gartner**	**IBM**
Taso 0	*kyvykkyyden puuttuminen*	*tietämätön:* Datalle ei ole määritetty omistajuuksia, turvatoimenpiteitä tai tapaa toimia sen suhteen.	*N/A*
Taso 1	*alustava tai tapauskohtainen:* Vähän tai ei ollenkaan hallinnointia. Rajallinen työkaluvalikoima. Siilokohtaiset roolit. Epäjohdonmukaisesti sovelletut kontrollipisteet tai ei kontrollipisteitä ollenkaan. Datan laadun ongelmia ei käsitellä.	*tietoinen:* Liiketoiminta- ja IT-johtajat alkavat ymmärtää ja tiedostaa tiedon ja organisaation tiedonhallinnan arvon.	*alustava:* Ei yhtään tai vähän ymmärrystä datan tärkeyden ymmärryksestä. Ei asetettuja standardeja datan hallinnoimiseksi.
Taso 2	*toistettavissa:* Kehittyvä hallinnointi ja osittain yhdenmukaiset työkalut. Joitain määriteltyjä rooleja ja prosesseja. Kasvava tietoisuus datan laadun ongelmien vaikutuksesta.	*reaktiivinen:* Tiimit jakavat tietoa keskenään. Tiedonhallinnan mukaan toiminen on vähäistä.	*hallittu:* Datan tärkeys organisaatiossa on ymmärretty.
Taso 3	*määritelty:* Data nähdään organisatorisena mahdollistajana. Yhdenmukaiset ja skaalattavat prosessit ja työkalut. Vähemmän manuaalisia vaiheita. Prosessien tuotokset ovat ennustettavampia.	*ennakoiva:* Tiedonhallinnan mukainen toiminta hyväksytään ja otetaan käyttöön. Tiedon hallintamalli tulee osaksi jokaista projektia.	*määritelty:* Datan regulointi ja hallinnoinnin ohjeet on määritelty paremmin ja ne on integroitu organisaation prosesseihin.
Taso 4	*hallittu:* Keskitetty suunnittelu ja hallinnointi. Dataan liittyvien riskien hallinta. Datanhallinnan metriikat. Mitattavaa datan laadun kehitystä.	*hallittu:* Tiedonhallinnan standardit ja käytännöt ovat hyvin ymmärrettyjä ja jalkautettuja.	*määrällisesti hallittu:* Määrälliset tavoitteet on asetettu jokaiselle projektille, dataprosessille ja ylläpidolle.
Taso 5	*optimoitu:* Ennustettavat prosessit. Alentunut riskitaso. Hyvin ymmärretyt metriikat datan laadun ja prosessien laadun hallintaan.	*tehokas:* Organisaatio on saavuttanut tiedonhallinnan tavoitteensa.	*optimoidaan:* Tiedon hallintamallista tulee organisaation laajuinen, mikä parantaa tuottavuutta ja tehokkuutta.

Maturiteettitasolta yksi lähtien datan merkitys ymmärretään organisaatiossa jo jollain laajuudella ja vakavuusasteella. Tällä maturiteettitasolla organisaation johdossa on herätty tiedostamaan datan ja datanhallinnan arvo liiketoiminnalle, mutta datanhallinnan periaatteiden mukaan toimiminen on kuitenkin vielä enemmän tiedostamatonta, siiloutunutta ja vailla ohjaavia käytäntöjä ja standardeja. DAMA:n ja IBM:n maturiteettimallit korostavat ensimmäisen maturiteettitason otsikoinnilla datanhallinnan alustavuutta, kun taas Gartner korostaa datanhallinnan tietoisuuden kasvua olemattomasta näkyväksi.

Maturiteettitasolla kaksi ymmärrys datan arvosta on levinnyt jo laajemmalle ja kehitettyjen käytäntöjen ja standardien mukainen toiminta on toistettavissa muilla liiketoiminta-

alueilla. Toiminta on kuitenkin vielä pitkälti reaktiivista. IBM:n eroaa tämän maturiteettitason otsikoinnilla muista malleista ja kutsuu jo tätä tasoa hallituksi datanhallinnaksi. Gartner korostaa toiminnan reaktiivisuutta ja DAMA toiminnan toistettavuutta.

Maturiteettitasolla kolme datanhallinta on siirtynyt reaktiivisesta kohti keskitetysti ohjattua ennakoivaa toimintaa, jonka seurauksena datan laadussa ja prosesseissa huomataan positiivista kehitystä. DAMA ja IBM korostavat, että tällä tasolla datanhallinnan toiminta on määriteltyä, kun taas Gartner korostaa datanhallinnan ennakoitavuutta.

Maturiteettitasolla neljä organisaation datavarannot ovat keskitetysti omistettuja ja hallittuja. Datalle ja datanhallinnan osa-alueille on asetettu mittareita, joilla seurataan liiketoiminnan niille asettamia tavoitteita. IBM:n datanhallinnan maturiteettimalli korostaa, että tällä tasolla datanhallinta on määrällisesti hallittua, kun DAMA ja Gartner pitävät tätä ensimmäisenä varsinaisena hallitun datanhallinnan maturiteetin tasona.

Korkeimmalla eli datanhallinnan maturiteettitasolla viisi datan ja datanhallinnan kehitys on jatkuvaa ja kehittämisen hyödyt näkyvät laajasti organisaation toiminnassa ja tuloksessa. Lisäksi datanhallintaa optimoidaan aina organisaation uusien tavoitteiden ja strategian mukaan. Kaikki maturiteettimallit korostavat korkeimman tason osalta datanhallinnan optimoinnin kautta saatavaa toiminnan tehokkuutta.

DAMA:n julkaiseman DMBOK-kirjan tarkoitus on antaa kattava, standardoitu kokonaiskuvaus datanhallinnan osa-alueista ilman kytköstä tiettyihin metodeihin ja tekniikoihin. Maturiteettianalyysi voidaan kehittää organisaation tarpeita vastaaviksi valiten analyysiin joko kaikki osa-alueet tai painottaen vain tiettyjä osa-alueita. (Roe 2011). DAMA:n datanhallinnan osa-aluekohtaista kriteeristöä vasten voidaan asettaa kysymyksiä sekä datanhallinnan aktiviteettien, standardien, työkalujen että henkilöresurssien maturiteetista analyysia varten. (Sebastian-Coleman 2020). Loihde Advisory Oy:n datanhallinnan asiantuntijat ovat laajasti omaksuneet DAMA:n datanhallinnan osa-aluejaon ja soveltavat DAMA:n ohjeistuksia käytännössä. Osa asiantuntijoista on suorittanut DAMA:n tarjoaman sertifioinnin datanhallinnan ammattilaiseksi. Niinpä on luonnollista, että myös tekoälykohtaisen datanhallinnan maturiteettianalyysi mukailee asiantuntijoille jo tutuksi tullutta DAMA-pyörän tietoaluejakoa.

Kun maturiteettianalyysi on suoritettu, analyysin tulokset voidaan esittää esimerkiksi tutkakaaviossa, tutummin hämähäkkikaaviossa, johon voidaan sijoittaa sekä nykytilan että tavoitetilan mukaiset maturiteettiasteikot datanhallinnan osa-alueittain. Kaaviota voidaan hyödyntää myös osoittamaan tapahtunutta kehitystä eri arviointien välillä.

(Sebastian-Coleman 2020). Tässä opinnäytetyössä tutkakaaviotyyppiä on hyödynnetty visualisoimaan opinnäytetyön tuloksia 5.1-kappaleessa.

2.4 Tekoälykohtaisen datanhallinnan maturiteetin arviointi

Monet organisaatiot havittelevat pääsevänsä hyötymään edistyneestä teknologiasta ja käytännöistä muun muassa data-analytiikan osa-alueella. Jotta edistyksestä on mahdollista liiketoimintahyötyä, datanhallinnan perustusten on oltava riittävällä tasolla tukeakseen edistyneiden käytänteiden jalkauttamista. Liiketoimintatiedon hyödyntäminen ja analytiikka ovat riippuvaisia kaikista muista datanhallinnan osa-alueista joko suoraan tai epäsuorasti. Sue Geuensin kehittämän mallin mukaan perusta liiketoimintatiedon hyödyntämiselle ja analytiikalle lähtee datan hallinnoinnista, johon sisältyy niin metadatan hallinta, tietoturva, data-arkkitehtuuri kuin viitetiedon hallintakin. Kaikki muut datanhallinnan osa-alueet ovat riippuvaisia näistä. Tämän pohjan päälle voidaan rakentaa luotettavan datan laadun, datan suunnittelun ja datan yhteentoimivuuden varmistavia käytäntöjä datan tallennus ja toiminnot -osa-alueelle. Edellisten maturiteetti näkyy järjestelmien ja sovellusten luotettavuutena, josta päästään taas korkeampaan ydintiedon hallinnan ja datavarastojen maturiteettitasoon. Kun nämäkin datanhallinnan osa-alueet ovat riittävällä maturiteettitasolla, tavoitetaan edistyneen liiketoimintatiedon hyödyntämisen ja analytiikan täysimääräisen potentiaalin hyödyntämismahdollisuudet. (Technics Publications 2017, 41–42).

Kun organisaatiot siirtyvät tekoälyn aikakaudelle, jokainen datanhallinnan osa-alue tulee ottaa kehityksessä huomioon ja saattaa riittävälle maturiteettitasolle, jotta saavutetaan tuotantokelpoinen ja pieniriskinen tekoälyratkaisu. Niinpä myös tekoälykohtaisen datanhallinnan maturiteetin arvioinnissa on analysoitava kaikki datanhallinnan osa-alueet sillä liiketoiminta-alueella, jota tekoälykehitys koskee. Tässä opinnäytetyössä tutkitaan, mikä on riittävä datanhallinnan maturiteettitaso AI-valmiille, tekoälykehityksessä tuotanto- ja elinkaarivaiheeseen siirtyvälle organisaatiolle. Jos organisaatio on vasta aloittamassa AI-matkaansa, opinnäytetyö antaa myös suuntaviivoja datanhallinnan kehityksen priorisointia varten.

Viime vuosina eri maiden hallitukset ja kansainväliset organisaatiot ovat laajasti heränneet kehittyvän AI-teknologian maailmaan ja kehittäneet periaatteita ja suosituksia tekoälyn hallinnointia varten. Yhtenä esimerkkinä tällaisesta suosituskehyksestä on yhden johtavan AI-maan Singaporen maailman talousfoorumissa vuonna 2019 julkaisema AI:n hallintamalli – Model AI Governance Framework, jossa annetaan käytännön suosituksia

tekoälyn käyttöönottoon. Malli painottaa erityisesti kriittisyyttä kehittää ihmislähtöisiä tekoälyratkaisuja, joissa tekoälyavusteinen päätöksentekoprosessi on selitettävä, läpinäkyvä ja oikeudenmukainen. (PDPC 2020).

Singaporen AI:n hallintamallista on identifioitavissa suosituksia myös datanhallinnan osa-alueille. Datan hallinnoinnin osalta malli suosittelee sellaisen hallinnointirakenteen luomista, jossa ymmärretään tekoälypohjaisen päätöksenteon arvot, riskit ja vastuut. Selkeitä, eri tasoisia rooleja ja vastuita tarvitaan tekoälykehityksen jokaisessa vaiheessa monitoroimaan, hallinnoimaan ja minimoimaan riskejä sekä kouluttamaan henkilöstöä tekoälyn käyttöönottoon ja hyödyntämiseen liittyvistä uusista vaadittavista käytänteistä. Lisäksi jokaisella tekoälykehityksessä mukana olevalla liiketoiminta-alueella täytyy olla sen alueen datan laadusta vastaavat roolit ja henkilöt. (PDPC 2020). Datan hallinnoinnin ja tietoturvan on siis oltava maturiteetiltaan kehittynyttä jo tekoälykehityksen alkuvaiheessa, jotta datan arvo ja riskit ylipäätään ymmärretään.

Tekoälyratkaisua suunniteltaessa tulee ottaa huomioon kaikki ratkaisun käyttötarkoitukset. Tämän ymmärryksen kautta voidaan kuvata tekoälyratkaisun tiedonkulkuarkkitehtuuri. Arkkitehdit varmistavat, että tekoälymallit ovat kestäviä ennen mallien käyttöönottoa. (PDPC 2020). Tämä edellyttää, että organisaatiossa on oltava data-arkkitehtuuritoiminto olemassa ja että arkkitehtuurikäytännöt ovat jo jollain asteella vakiintuneet.

Tietoturvan ja tietosuojan osalta Singaporen hallintamallissa suositellaan hyödyntämään olemassa olevia riskinhallinnan ja riskikontrollin toimenpiteitä analysoimaan ja hallinnoimaan tekoälyn jalkauttamisen riskejä sekä yleisesti organisaatiolle ja sen liiketoiminnalle että yksilöille, joihin tekoäly mahdollisesti vaikuttaa. Yksilöjen osalta on arvioitava muun muassa se, pitääkö heille tarjota vaihtoehto kieltäytyä – opt-out, tekoälyratkaisun käytöstä. Sellaisten tekoälyratkaisujen osalta, joiden toimintaan yksilö voi vaikuttaa vahingollisesti syöttämällä dataa manipulointitarkoituksessa, tarvitaan lisäyksiä käyttöehtoihin, jotta tällainen toiminta minimoidaan. Riskitaso ja riskinsieto määritetään sen mukaisesti, millaista tekoälyratkaisua organisaatio on kehittämässä. Esimerkiksi tuotteita tai lääketieteellistä diagnoosia ehdottavalle tekoälyratkaisulle määritetään keskenään hyvin erilaiset riskitasot. Kaikkien tekoälyratkaisujen tulee kuitenkin olla auditoitavissa niin algoritmin, datan kuin suunnitteluprosessinkin osalta. Auditoitavuutta voidaan tukea erilaisin keinoin, kuten esimerkiksi logitusten ja kattavan dokumentaation kautta. (PDPC 2020). Organisaatiossa tulee olla tietoturvan ja tietosuojan osalta olemassa olevia ja toimivaksi havaittuja hallintakeinoja olemassa, joiden päälle rakennetaan lisäksi nimenomaan tekoälyn riskejä kontrolloivia ja tekoälyn auditoitavuutta mahdollistavia keinoja.

Tekoälymallien kehittämisessä voidaan hyödyntää hyvin moninaisia, niin ulkoisia kuin sisäisiäkin datalähteitä. Hyödynnetyn datan elinkaari on tunnettava hyvin eli mikä on datan alkuperäinen lähde, miten sitä kerätään, käytetään, muokataan, rikastetaan ja jo hyödynnetään organisaatiossa sekä miten datan laatua hallitaan. Tekoälyratkaisun monitorointiin voidaan rakentaa esimerkiksi niin kutsuttu musta laatikko, joka tallentaa kaikki saapuvat datavirrat ja tallentaa tapahtumat ja tekniset ongelmat. Lisäksi elinkaaren dokumentointi auttaa datavinoumien jäljittämisessä ja korjaamisessa. (PDPC 2020). Datanhallinnan osa-alueista siis myös datan tallennus ja toiminnot -aktiviteettien maturiteetin on oltava määritellyllä tasolla, jotta mahdollistetaan tekoälyn virheellisen toiminnan tehokas korjaaminen, kun ongelmat johtuvat datan sisällöstä ja laadusta.

Organisaation on tehtävä toimenpiteitä, joilla varmistetaan kaiken tekoälykehityksessä hyödynnetyn datan tarkoituksenmukaisuus. Tätä varten data on ymmärrettävä riittävällä tasolla, jotta tiedostetaan muun muassa se, millä eri tavoin data voi vinoutua. Toisin sanoen riittävän datalähtöisen ymmärryksen kautta voidaan minimoida vinoumat ja vähentää niistä koituvia riskejä. Vinoumat voivat johtua esimerkiksi siitä, että data ei kata kaikkia mahdollisia skenaarioita, dataa kerätään virheellisesti tai uutta dataa lisätään jo tekoälyllä testattuun datajoukkoon, jolloin tekoälymallin tulokset painottuvat väärin. Tekoälyn näkökulmasta datan laatua täytyy arvioida hyvin moninaisilla mittareilla. Näitä mittareita ovat datan täsmällisyys, täydellisyys, totuudenmukaisuus, ajantasaisuus, merkityksellisyys, eheys, käytettävyys ja se, kuinka paljon ihminen on muokannut dataa. Mitä tarkoituksenmukaisempaa tekoälykehityksessä hyödynnettävä data on ja mitä enemmän sitä on, sitä tarkempia ja oivaltavampia tekoälymallitkin ovat. Tekoälymallit sekä niiden hyödyntämä data ja kehitysvaiheet on myös dokumentoitava kattavasti, jo pelkästään kriittisten tekoälykehitykseen liittyvien roolien henkilöstövaihdosten varalta. (PDPC 2020). Datan laatu ja metadatanhallinta ovat siten riippuvaisia toisistaan, sillä ilman ymmärrystä datasta ei ole myöskään ymmärrystä datan laadusta eikä kyetä tekemään datan laatua korjaavia toimenpiteitä, jotka ovat kriittisiä tekoälyratkaisun kehittämisessä. Datasta riippuen tulee maturiteettivaatimuksia myös sille datalle, jota tekoälyratkaisu tulee hyödyntämään, olkoon data esimerkiksi viite- tai ydintietoa tai strukturoimatonta dataa. Myös dokumenttienhallinta on oltava riittävällä maturiteettitasolla, jotta dokumentointikäytännöt ovat standardoituja ja dokumentit keskitetysti ja helposti saatavilla koko organisaatiossa.

Tekoälykehityksessä mukana olevien analytiikka-asiantuntijoiden täytyy osata tulkita tekoälymallien tuotoksia ja päätöksiä, jotta mahdolliset vinoumat havaitaan nopeasti. Monitoroinnin tukena voidaan hyödyntää myös automatiikkaa. Lisäksi tarvitaan prosessi, jonka kautta myös muut tekoälyä hyödyntävät sidosryhmät voivat raportoida havaitsemistaan epäkohdista oikeille tahoille. (PDPC 2020). Organisaation analytiikkakyvykkyyden on

siis oltava jo rakennettuna prosessien ja tarvittavien kompetenssien osalta, jotta sitä voidaan laajentaa kohti tekoälyn hyödyntämistä.

Singaporen AI:n hallintamallin ohjeistuksista on identifioitavissa suurin osa datanhallinnan osa-alueista. Lisäksi identifioiduille osa-alueille kohdistuu vähimmäismaturiteettivaatimuksia, jotta tekoälykehitys ja kehitettävä tekoälyratkaisu olisivat datanhallinnan osalta sekä vaatimusten- ja lainmukaisia että liiketoiminnalle hyödyllisiä. Tämän opinnäytetyön tarkoitus on tarkemmin tutkia, mitä vähimmäismaturiteettia vaaditaan kullekin datanhallinnan osa-alueelle sekä tekoälykehitykseen lähdettäessä että kun tekoälyratkaisu viedään tuotantoon.

3 Tutkimus- ja kehittämismenetelmät

Tässä osiossa käydään läpi tämän opinnäytetyön kehittämisosion lähestymistapa, vaiheet ja menetelmät. Opinnäytetyö suoritettiin toimeksiantotyönä Loihde Advisory Oy:lle, joka muutti nimensä opinnäytetyön kirjoittamisen aikana Talent Base Oy:sta Loihde-konserniin liittymisen myötä. Opinnäytetyössä kehitettiin uusi toimintamalli asiakashankkeiden datanhallinnan alkuanalysointivaiheeseen, jossa on otettava huomioon se, että aiempaa useampi asiakasorganisaatio harkitsee hyödyntävänsä tai jo hyödyntää tekoälyä.

3.1 Lähestymistapa

Opinnäytetyön kehittämisosiossa käytettiin Ojasalon, Moilasen ja Ritalahden (2015, 37–66) kirjassaan esittelemää konstruktiivista, tutkimuksen hyödyntäjän ja toteuttajan väliseen kommunikointiin perustuvaa lähestymistapaa, jota voidaan soveltaa silloin, kun rakennetaan teoriaan pohjautuvaa ratkaisua. Opinnäytetyön tavoitteena oli tuottaa konkreettinen tuotos eli tässä tapauksessa tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli. Maturiteettimalli on tarkoitettu käytännön työhön osaksi datanhallinnan konsulttien työkalupakkia. Painotetun datanhallinnan maturiteettimallin kehittäminen sidottiin aikaisempaan teoriaan ja sen, sekä haastattelujen ja ideatyöpajan kautta toteutettiin ratkaisu, jonka hyödyllisyyttä testataan jo sovitussa asiakasprojektissa opinnäytetyön jälkeen. Hyödyllisyys näkyy siinä, miten organisaatiot kokevat saavansa maturiteettianalyysin tuloksena konkreettisia askelmerkkejä kohti AI-valmista organisaatiota ja kuinka paljon konsultit tulevat hyödyntämään maturiteettimallia arvioidessaan organisaatioiden datanhallinnan kyvykkyyttä tekoälykehityksessä.

3.2 Aineiston hankintamenetelmät

Tämä opinnäytetyö on luonteeltaan kvalitatiivinen. Työn kehittämisosio oli uudistamisperustainen eli perinteisille organisaatioille kohdistettua datanhallinnan maturiteettimallia uudistettiin painottamalla tekoälykehityksen kannalta kriittisiä datanhallinnan osa-alueita. Uudistamisperustaisella kehittämistyöllä tarkoitetaan sitä, että esimerkiksi uusi malli tai toimintaprosessi kehitetään eri näkökulmia haravoimalla ja sitä kautta uutta ideoimalla (Ojasalo ym. 2015). Kehittämistehtävän aineiston keruu toteutettiin haastattelemalla pääasiassa Loihde Advisory Oy:n datanhallinnan asiantuntijoita (liite 2), joilla on vuosien, jopa vuosikymmenten kokemuspohja erilaisista datanhallinnan asiakashankkeista useilla eri toimialoilla. Lisäksi haastateltiin Loihde Advisory:n ulkopuolella työskenteleviä Loihde-konsernin työntekijöitä, joilla on käytännön kokemusta tekoälyratkaisujen

suunnittelusta ja rakentamisesta. Haastatteluilla saatiin siis kerättyä monipuolisesti erilaisia näkökulmia, jotka tukevat tämän organisaation työntekijöille kohdennetun työkalun eli tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin rakentamista.

Haastattelumäärä rajattiin noin kymmeneen haastatteluun tai kun saturaatio olisi tullut täyteen eli kun ei enää olisi saatu kyseisellä menetelmällä uutta informaatiota. Lopulta opinnäytetyötä varten haastateltiin yhtätoista Loihde-konsernin datanhallinnan ja tekoälyn asiantuntijaa etänä Teams-sovelluksen kautta syksyllä 2021. Asiantuntijat valittiin yhdessä työpaikkaohjaajien kanssa. Kaikki haastateltavat ovat potentiaalisia opinnäytetyön tuotoksen eli tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin hyödyntäjiä, joten luonnollisesti heidän näkemyksensä toivat tähän työhön arvoa.

Haastattelut kestivät noin tunnin. Haastattelut nauhoitettiin käyttäen Teams-sovelluksen nauhoitustoimintoa. Tämän jälkeen aineisto litteroitiin ensin Word-selainversion litterointitoiminnon avulla, jonka jälkeen teksti puhtaaksikirjoitettiin kuunnellen samalla nauhoitusta. Litteroinnit validoitiin lähettämällä puhtaaksikirjoitettu teksti haastateltaville sähköpostitse ja pyytämällä kuittaus vastausten tarkistamisesta. Haastatteluaineiston litteroinnissa keskityttiin kirjaamaan ylös asiasisältöä sisältävät lauseet ilman täytesanoja ja maneereja, koska opinnäytetyön luonteen vuoksi viimeksi mainituilla ei nähty olevan arvoa tässä asiayhteydessä.

Asiantuntijoiden haastattelukysymykset koostettiin mukailemaan opinnäytetyön teoriapohjan aihealueita. Haastateltaville kerrottiin, että vastauksia käytetään datanhallinnan maturiteettimallin luontiin, joten on tärkeää, että he tuovat esille niitä asioita, joita heidän mielestänsä pitäisi selvittää datanhallinnan osalta organisaatioissa, joissa halutaan hyödyntää tekoälyä. Haastattelut olivat puolistrukturoituja eli ennalta laadittujen kysymysten lisäksi voitiin lisäksi kysyä haastattelun kuluessa mieleen tulevia uusia relevantteja kysymyksiä ja toisaalta jättää muutamissa haastatteluissa tietyt kysymykset vähemmälle huomiolle, jos kyseisellä haastateltavalla ei ollut kyseessä olevaan asiaan kokemuspohjaa. Haastatteluissa käytiin läpi datanhallinnan ja tekoälykehityksen yhdistäviä teemoja. Kysymykset laadittiin sen mukaan, millaista tietoa tarvittiin kehittämistyön tueksi. Runko mukaili opinnäytetyön teoriapohjaa.

Haastattelujen ja teoriapohjan mukaan kehitettyä tekoälykehityksen mukaan painotettua datanhallinnan maturiteettimallia validoitiin ja jatkojalostettiin neljän hengen yhteisöllisessä ideointityöpajassa AIGA-hankkeessa mukana olevien Loihde Advisory -konsulttien kesken. Ideointityöpajan aikana validoitiin haastatteluissa korostettujen datanhallinnan

osa-alueiden painopisteitä tekoälykehitysvaiheittain. Kerätyt lisähuomiot on sisällytetty aineiston analyysiin vastaaville aihealuekohdille 4.3-kappaleessa.

Ideointityöpajassa olivat edustettuna idearikkaat asiantuntijat sekä tekoälyn, datanhallinnan että myynnin puolelta. Näin varmistettiin, että maturiteettianalyysimalli ottaa sekä datanhallinnan että tekoälyasiantuntijoiden intressit huomioon, jotta malli ei aseta liian suuria datanhallinnan muureja tekoälykehitysmatkalle lähtemiseen, mutta ei toisaalta myöskään kannusta siilomaisiin tekoälykokeiluihin, joita ei olla linkitetty organisaation strategiaan ja jotka eivät ole skaalattavissa. Lisäksi myynnin edustus varmisti sitä, että maturiteettimalli toimii asiakasorganisaatioille myytävänä kokonaisuutena.

Työpajassa hyödynnettiin verkossa toimivaa ideointityökalu Miroa, jonka digitaaliselle valkotaululle rakennettiin haastatteluaineiston perusteella tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli. Datanhallinnan osa-aluekohtaiset maturiteettitasovaatimukset per tekoälykehitysvaihe ilmoitettiin keskiarvona haastatteluissa kerätyistä arvoista. Lisäksi maturiteettimalliin sisällytettiin kattavasti lainauksia haastatteluaineistosta ideoinnin tueksi.

Ideointityöpajan aluksi käytiin läpi aivoriihen tavoitteet ja toimintaperiaatteet muun muassa kommentoinnin suhteen. Ideointivaihe jaettiin kahteen äänestysosioon, jota seurasi kommentointikierros. Äänestyksessä haettiin validointia haastatteluissa korostettujen datanhallinnan osa-alueiden ja tekoälykehitysvaiheiden välisten suhteiden painottamisen oikeellisuudelle. Taulukkomuotoisessa maturiteettimallissa (kuva 5, 27) oli haastattelujen perusteella korostettuna vaaleanvihreällä ne tekoälykehityksen vaiheet, joihin mennessä kukin datanhallinnan osa-alue tulisi kehittää taulukon kyseiseen laatikkoon kirjoitetulle maturiteettitasolle. Sinertävällä värillä oli korostettu ne tekoälykehityksen vaiheet, joihin mennessä datanhallinnan osa-alueita tulisi jatkokehittää ylemmälle maturiteettiasteelle.

Ideointityöpajan kaksivaiheisessa äänestyksessä osallistujia pyydettiin merkitsemään vaaleanvihreällä taustavärillä korostetuista laatikoista vihreällä pallolla datanhallinnan osa-alueista ne, joiden painotus tiettyyn tekoälykehitysvaiheeseen mennessä vaikutti paikkansapitävältä. Keltaisella pyydettiin merkitsemään epäselvät tai epävarmat kohdat. Punaisella pallolla pyydettiin merkitsemään datanhallinnan osa-alueista ne, joiden painotus nähtiin olevan merkitty väärään tekoälykehitysvaiheeseen mennessä. Kummankin

äänestysvaiheen jälkeen keltaisella ja punaisella pallolla merkityt kohdat käytiin läpi kommenttikierroksen kautta.

Kuva 5. Työpaja tekoälykehityksen mukaan painotetusta datanhallinnan maturiteettimallista

3.3 Aineiston analyysimenetelmät

Tämän opinnäytetyön haastatteluaineiston analyysi oli teoriaohjauksinen. Tämä tarkoittaa, että aineistoa tarkastellaan teoriapohjasta lähtöisin olevien valmiiden teorioiden ja käsitteiden kautta (Ojasalo ym. 2015, 140). Tällä menetelmällä varmistettiin, että lopullinen tuotos eli tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli on riittävän kattava ja validi datanhallinnan maturiteetin analysoinnin väline organisaatioissa, joissa halutaan upottaa tekoäly osaksi liiketoiminnan operatiivista toimintaa. Koska etukäteen tiedostettiin, että aineistosta voi nousta esiin yllättäviäkin teemoja, jotka on syytä ottaa huomioon tekoälyä hyödyntävissä organisaatioissa, aineiston analyysia ei

haluttu sitoa liian tiukasti teoriaan. Lisäksi tietoperustan kirjoittamisen aikana hyödynnettiin oivalluttava–perinteinen -mallia. Tässä mallissa referoinnin ja haastattelulainausten sekaan lisätään kehittämistyöhön liittyvää omaa ajattelua, mutta tulokset sijoitetaan omaan osioonsa (Ojasalo ym. 2015, 35).

Haastatteluaineistojen valmistelun, tässä tapauksessa litteroinnin jälkeen aineiston analyysia varten luotiin teoriapohjan mukainen strukturoitu analyysirunko. Tällaisella rungolla voidaan tutkia teoriataustan toimivuutta uudessa kontekstissa ja identifioida sekä analyysirungon sisälle että ulkopuolelle kuuluvat teemat (Ojasalo ym. 2015, 140). Strukturoitu analyysirunko tuki tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin kehittämistä jo olemassa olevien datanhallinnan maturiteetti- ja AI-hallintomallien pohjalta. Tekstimuotoinen haastatteluaineisto merkittiin teemojen mukaisesti värikoodein, jotta analyysivaiheessa löydettiin helposti käsiteltävään teemaan liittyvät vastaukset. Aineistosta etsittiin haastattelujen välisiä säännönmukaisuuksia ja poikkeavuuksia. Näitä havaintoja tarkasteltiin suhteessa läpikäytyyn teoriaan ja pyrittiin luomaan kokonaisvaltainen pohja opinnäytetyön tuotokselle.

Haastatteluvastauksista oli selkeästi poimittavissa keskenään linjassa olevia datanhallinnan osa-alueiden painotuksia suhteessa tekoälykehitykseen ja sen eri vaiheisiin. Teoriapohjan ja haastattelujen analysoinnin perusteella rakennettiin painotettu datanhallinnan maturiteettimalli tekoälykehitykseen lähtevälle tai jo lähteneelle organisaatiolle. Suurin osa haastateltavista tunsi tai ymmärsi DAMA:n datanhallinnan osa-aluejaottelun. Lisäksi haastateltavat näkivät, että analysoitaessa tekoälykehitykseen lähtevän organisaation datanhallinnan maturiteettia analyysin teemat voivat hyvin olla samoja, kuin perinteisissä datanhallinnan maturiteettimalleissa, mutta vain eri painotuksilla. Tämän perusteella DAMA:n datanhallinnan osa-alueet ja niiden sisältämä kriteeristö valittiin tämän työn pohjaksi ja tuotosta täydennettiin haastattelujen havainnoilla ottamaan huomioon tekoälykehityksen erityispiirteet.

Kuten aiemmin 3.2-kappaleessa on kerrottu, alustavaa versiota tekoälykehityksen mukaan painotetusta datanhallinnan maturiteettimallista käytiin läpi ideointityöpajassa. Työpajan äänestyksen jälkeen muistiinpanot kirjattiin ylös ja muutokset merkittiin taulukkoon korostusväreillä. Kommenttien perusteella punaisella värillä korostettiin ne tekoälykehitysvaiheet datanhallinnan osa-alueittain, joita työpajaan osallistuneet pitivät joko lisäksi tai toisen vaihtoehdon sijasta oikeampana datanhallinnan kehityksen tai jatkokehityksen painopisteenä.

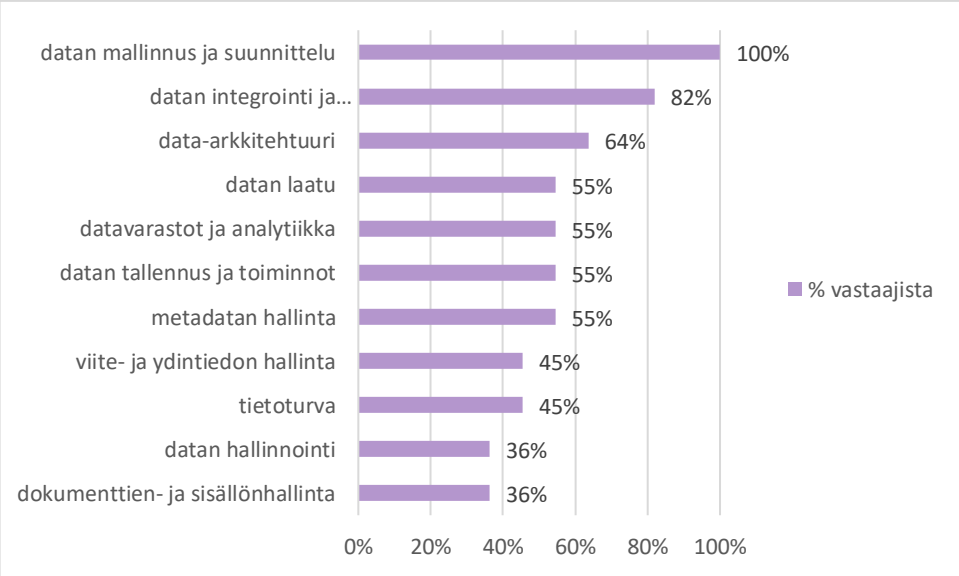
Ideointityöpajan jälkeen vuorossa oli tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin viimeistely, jossa hyödynnettiin benchmarking-menetelmää. Tässä menetelmässä etsitään esimerkiksi markkinoilla jo olevia hyviä malleja ja sovelletaan niistä kerättyjä havaintoja ja oppeja oman mallin kehittämiseen (Ojasalo ym. 2015, 186). Painotettua maturiteettimallia peilattiin Singaporen AI-hallintakehyksen mallia vasten, koska Singapore on yksi johtavista tekoälyvaltioista. Näin varmistettiin, että painotettu datanhallinnan maturiteettimalli ottaa kantaa myös Singaporen AI-hallintakehyksen sisältämiin vaatimuksiin datanhallinnan osalta, mikä varmistaa maturiteettimallin laadun.

4 Aineiston analyysi

Opinnäytetyön tutkimusosuudessa kartoitettiin haastattelujen ja ideointityöpajan kautta Loihde-konsernin datanhallinnan ja tekoälyn asiantuntijoiden näkemyksiä datanhallinnan niistä osa-alueista, joiden kehittämiseen heidän mielestään tulisi panostaa tekoälykehityksen eri vaiheisiin mennessä, jotta voidaan saavuttaa datanhallinnan osalta tuotantokelpoinen ja skaalattavissa oleva tekoälyratkaisu.

Haastattelukysymykset (liite 1) jakaantuivat yleisiin kysymyksiin, kysymyksiin hyvästä datanhallinnasta sekä kysymyksiin datanhallinnan merkityksestä tekoälykehityksessä. Haastattelun ensimmäiset yleiset kysymykset olivat kategorisoivia kysymyksiä, jolla kartoitettiin millä datanhallinnan osa-alueilla (kuvio 1) ja tekoälykehityksen vaiheissa (kuvio 2, 31) haastateltavat ovat työskennelleet.

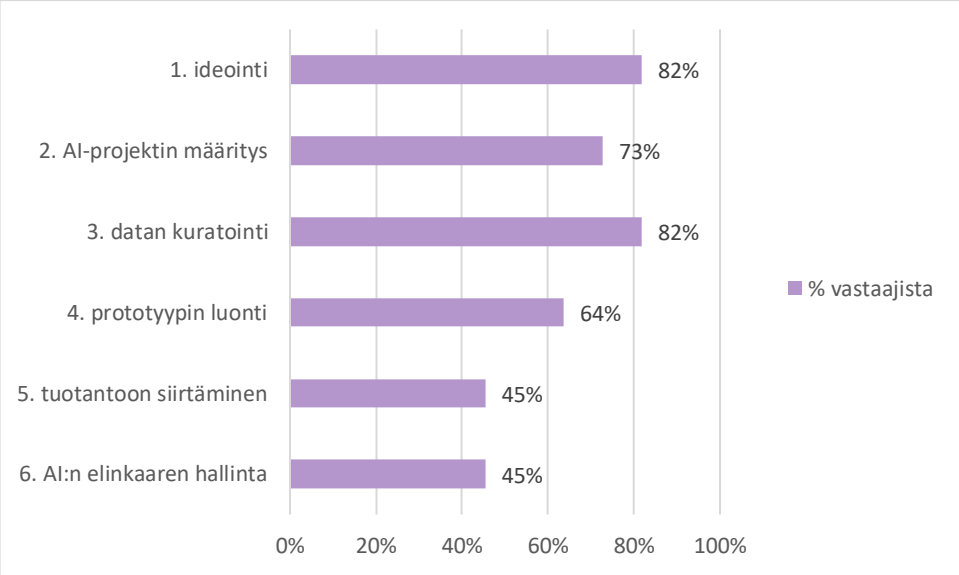
Jokainen haastateltava on työskennellyt useammalla datanhallinnan osa-alueella. Kaikki haastateltavat ovat työskennelleet datan mallinnus- ja suunnittelupuolen tehtävissä, osa vuosia ja osa taasen silloin tällöin projektien yhteydessä. Datan hallinnointi -työstä sekä dokumenttien- ja sisällönhallinnasta kokemusta löytyi määrällisesti vähiten, mutta nämä osa-alueet maininneilla oli näistä vuosien vankka kokemus.



Kuvio 1. Vastaajien datanhallinnan työkokemus osa-alueittain

Haastateltavilta löytyi datanhallinnan lisäksi kokemusta kaikista tekoälykehityksen vaiheista. Vastaajista 82 %:lla oli kokemusta AI-projektin ideoinnista ja datan kuratoinnista. Lisäksi yli puolella vastaajista oli kokemusta AI-projektin määrityksestä ja

prototyypin luonnista. Vähiten kokemusta löytyi tekoälyratkaisun siirtämisestä tuotantoon sekä AI:n elinkaaren hallinnasta.



Kuvio 2. Vastaajien tekoälykehityksen työkokemus kehitysvaiheittain

Lisäksi yleisillä kysymyksillä kartoitettiin, mitkä datanhallinnan ja tekoälyn standardit, regulaatiot ja maturiteettimallit ovat haastateltaville tutuimpia. Kaikki vastaajat kertoivat tutustuneensa GDPR:ään työnsä kautta. Vastaajista 27 % mainitsi myös tutustuneensa Euroopan parlamentin ja neuvoston asetusehdotukseen tekoälyn harmonisoidusta sääntelystä – Artificial Intelligence Act. Lisäksi 18 % vastaajista kertoi tuntevansa finanssialakohtaisia säädöksiä. Vastauksissa esiintyi yksittäisinä mainintoina myös useita muita toimialakohtaisia standardeja ja säädöksiä. Kukaan haastateltava ei maininnut hyödyntäneensä tekoälyyn liittyviä maturiteettimalleja, mutta datanhallinnan maturiteettimalleista Gartnerin datan hallinnointi -mallia oli hyödyntänyt 36 % vastaajista. Lisäksi 36 % vastaajista oli olemassa olevia maturiteettimalleja hyödyntäen luonut omia versioita datanhallinnan maturiteettimalleista kulloiseenkin asiakasprojektiin parhaiten soveltuvaksi.

Yleisten kysymysten jälkeen haastatteluissa siirryttiin kysymyksiin hyvästä datanhallinnasta ja siitä, mitä se haastateltaville ylipäätään tarkoittaa, minkälaisena he näkevät hyvän datanhallinnan merkityksen tekoälykehityksessä ja siitä koituvat lyhyen ja pitkän aikavälin hyödyt organisaatioille sekä mitä erityisesti pitäisi ottaa huomioon niiden organisaatioiden datanhallinnan maturiteetin arvioinnissa, jotka harkitsevat tekoälyn hyödyntämistä tai jo hyödyntävät sitä. Näihin liittyviä havaintoja analysoidaan tarkemmin 4.1- ja 4.2-kappaleissa.

Haastattelujen viimeisessä kysymysosiossa käytiin läpi tekoälykehitys vaihe vaiheelta. Vaihekohtaisesti kysyttiin, mitkä datanhallinnan osa-alueet ovat haastateltavien mielestä kriittisiä kyseessä olevassa vaiheessa, miksi ja millainen datanhallinnan maturiteettitaso näillä mainituilla osa-alueilla pitäisi olla, jotta voidaan puhua datanhallinnan osalta AI-valmiista organisaatiosta ja on kannattavaa edetä tekoälykehityksessä eteenpäin. Tämän osion läpikäyntiin käytettiin puolet haastatteluajasta. Näitä vastauksia analysoidaan tarkemmin 4.3-kappaleessa.

4.1 Hyvä datanhallinta

Hyvä datanhallinta lähtee ylemmällä tasolla siitä, että organisaation on aluksi ymmärrettävä, millaisia datavarantoja heillä ylipäätään on ja mitä mahdollisuuksia nämä voivat liiketoiminnalle tuoda. *”Hyvä datanhallinta ilmenee kokonaisvaltaisena ymmärryksenä organisaatiossa, että mitä tietoa, missä järjestelmissä, minkä prosessien käytössä ja on roolit ja vastuut kirkkaana sille, että kenen kuuluu tai jos on jotain muutoksia, joita tarvitaan datasisältöihin tai vaikka tietosuojamielessä niihin henkilötietojen käyttötarkoituksiin, niin ketkä tekevät päätöksiä, minkä prosessien kautta viedään muutokset läpi, jotta voidaan varmistua siitä, että esimerkiksi regulaatiot, mukaan lukien tietosuoja, regulaatiovaatimukset täyttyy eli roolit ja prosessit sille, että pysyy ajantasainen näkyvyys siihen, että mitä dataa ja missä ja mihin käyttötarkoituksiin.”*

Kokonaisvaltainen ymmärrys käytettävissä olevista datavarannoista antaa pohjan datan monipuoliselle hyödyntämiselle. *”...organisaatio pystyy hyödyntämään sitä dataa monipuolisesti ja joustavasti, että voidaan keksiä uusia tapoja käyttää sitä.”* Hyvä datanhallinta kuitenkin edellyttää, että organisaation johto on antanut datalle sen ansaitsemaa arvostusta. *”...joku tulisi sanoo oikeasti, että meillä on hyvin hallittu data ja me ollaan ylpeitä siitä, että me saadaan siitä oikeasti merkittävää toiminnan tehokkuutta sillä, että me ei tehdä päällekkäisiä asioita datan kanssa, data ei tuota meille suunnattomia virheitä ja näin päin.”* Tällöin käytännön datanhallinnan toimien kautta datan arvostus kääntyy organisaation tarjoamien tuotteiden ja palveluiden paremmaksi laaduksi. *”Mun mielestä hyvä datahallinta ilmenee asiakkaille tai yrityksen asiakkaille tuotettavien palveluiden laatuna.”*

Yleisesti ottaen haastateltavat kokivat hyvän datanhallinnan ilmentyvän suoraviivaisempana työnä, jossa ei kohdata dataan liittyviä ongelmia. *”...kaikki periaatteessa suunnitellut hyödyt saavutetaan suoraviivaisesti. Että sitten siellä ei ole datan laatuongelmia, ei kohdata saatavuusongelmia. Löytyy aina henkilöt, jotka tietää asioista ja on vastuussa asioista. Dokumentaatio on kunnossa. Että se työ on sujuvaa.”* Hyvä datanhallinta tekee

siis itsensä näkymättömäksi, koska fokus voidaan silloin pitää varsinaisessa liiketoiminta-hyötyä tuottavassa työssä. Tällöin datalle asetetut liiketoimintavaatimukset täyttyvät. *”Se ilmenee helppona työnä. Siis siinä, että asiat sujuu. (...) ...meillä oikeasti on sitä faktapohjaista tietoa itsellemme tuotettavissa sen datan perusteella. Me luotetaan siihen dataan, meillä on se silloin, kun me tarvitaan ja se on juuri sitä dataa, mitä me tarvitaan ja se laatu on meille riittävä.”*

Hyvä datanhallinta perustuu liiketoimintatarpeen mukaisesti määriteltyyn datan laatuun ja siihen, että data myös on määritellyn laadun mukaista. Tämä ei tietenkään tapahdu itsekseen vaan tarvitaan ihmisiä, prosesseja ja lopulta myös oikein valjastettua teknologiaa tämän toteuttamiseksi. *”Kaikkihan lähtee siitä, että sulla on laadukasta dataa eli tavallaan kaikki datanhallinnan toimenpiteet tähtää siihen, että sinulla on hyvälaatuista tietoa hyödynnettävissä organisaatiossa, on se sitten operatiivisiin järjestelmiin, analytiikkaan tai vaikka tekoälyn hyödyntämiseen. Eli se datan laatu on mun mielestä se ykkösjuttu. Sit kun siitä lähtee purkaa sitä, mitä se tarkoittaa, niin sitten päästään näihin muihin osa-alueisiin. Sulla pitää olla säännöstöt olemassa, johon se datan laatu viittaa ja sitä hyödynnetään datan laadun valvomiseen ja sulla pitää olla omistajat asioilla, että asiat tapahtuu, on se sitten strategisemman tason omistajaa, henkilöitä, jotka määrittelee näitä säännöstöjä tai sitten niitä ihan operatiivisen, jotka käsittelee sitä tietoa, niin ne pitää olla kunnossa eli se people-puoli ja prosessipuoli. Sit sulla pitää olla näitä tukeva teknologia eli sulla pitää olla hyvää teknologista, erityyppisiä hyviä käytänteitä, on se sitten data quality:in, on se sitten yhteisiä vaikka master data -järjestelmiä, mitä kautta asioita hallinnoidaan ja välitetään ja hyvät integraatiot, jotta data välittyy. Eli se teknologia pitää myös valjastaa siihen. (...) ...ihmisillä, jotka tätä hommaa koordinoi ja hallinnoi, pitää olla näkemystä, että miten näitä asioita, mitkä kaikki asiat tähän liittyy ja miten niitä tehdään vaiheistetusti ja järkevästi.”*

Hyvä datanhallinta varmistaa, että liiketoiminnan kriittiseksi määritelty data on koko organisaatiossa tiedossa ja hyödynnettävissä, data on sisällöltään ja laadultaan oikeanlaista, dataa monitoroidaan ja organisaatiosta löytyvät prosessit, roolit ja vastuut vastaamaan dataan liittyvistä toiminnoista. *”Kaikista tärkeintä on mun mielestä se, että tiedetään mikä on oleellista dataa, että on priorisoitu, että mikä on tärkeätä dataa. Etenkin kun puhutaan tästä oleellisesta datasta, niin se mikä näyttäytyy käyttäjälle, vaikka luvut tai keskiarvot, niin ne pitää paikkansa, että se on paikkansapitävää. Ja siinä pitää sitten ymmärtää, että mimmoisia virhelähteitä siinä matkan varrella siitä pisteestä missä se data syntyy, syntyykö se sitten jonkun ihmisen syöttämänä sinne järjestelmään vai jonkun mittaustuloksen, sensorin tai muun kautta tai jostain web-sivusta klikkauksella. Ja hallinnan näkökulmasta se on niin, että olisi hyvä valvoa datan laatua, riippuen sen datan kriittisyydestä, niin informaalikin tapa tarkkailla laatua voi olla riittävä. Ja tässä on tärkeätä semmoinen*

kulttuuri, missä pidetään keskeisenä sitä, että data on oikein ja on jonkinlainen paikka, minne voi raportoida havainnoista tai epäilyksistä sen laadun suhteen ja sitten, että on resursseja organisaatiossa, jotka pystyy selvittää ja korjata asioita, mikäli tämä tulee tarpeelliseksi."

Haastateltavilta kysyttiin myös työn kautta saatuja havaintoja huonosta datanhallinnasta. Vastausten perusteella huono datanhallinta tarkoittaa sitä, että datasta ei ole saatavilla tietoa, dataan ei ole luottamista eikä data siten ole helposti hyödynnettävissäkään. Koko datapääoman potentiaali jätetään siis tässä tapauksessa hyödyntämättä. *"Sitä on sirpaleisesti niin sanotusti siellä täällä ja siinä menetetään potentiaaleja, esimerkiksi datan yhdistelyä tai että tehdään tuplatyötä, koska data on monessa eri paikassa tai tämmöisiä ongelmia saattaa siihen sirpaloitumiseen sisältyä."*

Jos eri liiketoiminta-alueet nimeävät dataa kukin omalla tavallaan eikä dataa ole kartoitettu, mallinnettu eikä yhteistä sanastoa ole luotu, kasvaa riski väärinymmärryksille sellaisissa dataan liittyvissä projekteissa, joissa hyödynnetään dataa läpi organisaation. *"Saattaa olla, että siellä on jotain vakiintuneita käytäntöjä, että vaikka tietty termi tietyssä kohtaa saattaa tarkoittaa eri asiaa kuin jossain muualla. Tollanen tekee jotenkin hankalaa siitä datan hyödyntämisestä tai sellaisemmasta geneerisemmästä hyödyntämistä. Joutuu päättelemään hirveästi."* Jos lisäksi data ei ole yhteentoimivaa ja palvelee vain esimerkiksi yhtä liiketoimintaprosessia, dataa ei ole mahdollistakaan hyödyntää laajasti. *"...ettei se ole fiksaantunutta, että tätä tiettyä dataa voidaan käyttää vain tavalla A."*

Huono datanhallinta johtaa pitkittyneisiin projekteihin, kasvaneisiin kustannuksiin ja jopa siihen, ettei projekteja voidakaan jatkaa. *"Kun sitä projektia menee puolitoista kuukautta ennen kun pääsee kiinni dataan. Ettei se mene siihen, et sit ihmetellään, että no missä tää data ja mitä tämä tarkoittaa."* Hyvällä datanhallinnalla riskit ja kustannukset pienenevät, kun yllättäviä löydöksiä ja niistä johtuvia välttämättömiä toimenpiteitä esimerkiksi datan saatavuuden ja laadun osalta ei tule, koska dataan on kiinnitetty asianmukaista huomiota jo ennen datan hyödyntämiseen liittyviä hankkeita.

4.2 Datanhallinnan rooli tekoälykehityksessä

Kun haastateltavilta oli kartoitettu heidän näkemystään hyvästä datanhallinnasta, haastattelukysymyksissä siirryttiin tarkastelemaan datanhallinnan roolia tekoälykehityksessä. Kuten muutkin digitaaliset hankkeet, erityisesti myös tekoälyhankkeet ovat riippuvaisia datasta, jolloin datanhallinnan onnistuminen näkyy suoraan kehityshankkeiden onnistumisena. *"...käytännössä kaikki digitaaliset hankkeet, kehityshankkeet mitkä on, niin jossain*

määrin niissä on joku liityntäriippuvuus dataan. Yksi iso asia on jo se, että silloin kun meillä on oikeesti datanhallinta, meillä on siis operatiivisen ketterä, oikeasti hyvin toimiva datanhallinta, niin silloin esimerkiksi kaikkien kehityshankkeiden osalta data tukee niitä kehityshankkeita eikä se ole aina se kehityshankkeiden vaikein asia, että päästään dataan ja mitä tapahtuu. Me saadaan oikeasti merkittävää kehitystehokkuutta tai kehityksen tehokkuutta. Me löydetään siellä yhteisiä asioita, joilla saadaan vielä vietyä kehitystehokkuutta eteenpäin." Datanhallinta antaa myös eväitä priorisoida hankkeita datan liiketoiminta-arvon kasvattamisen kannalta. *"Tavallaan sen datanhallinnan kautta voidaan siis priorisoida datan kehittämiseen ja datan arvon kasvattamiseen olevat hankkeet."* Vaikka useat organisaatiot tiedostavat tekoälyn merkittävyyden nyt ja tulevaisuudessa, monet eivät kuitenkaan ymmärrä antaa tekoälyyn vahvasti kytköksissä olevalle datalle ja datanhallinnalle samanlaista arvoa. *"Mun mielestä tässä tekoälyn kohdalla on ehkä se, että ei ole vielä oikein herätty siihen, että kuinka merkittävä asia se data on."*

Tekoälykehityksessä on tärkeää, että muun muassa tarvittava data on ylipäätään saatavilla, identiteetin- ja pääsynhallintaprosessit ovat sujuvia ja datan yhteentoimivuus on taattu. *"Sitten on taas projekteja tyyliin ne luo käyttäjän johonkin ja sille käyttäjälle lisätään vaan oikeus, että saan nähdä jonkun tietyn datan ja sitten saatan saada sen päivässä. Henkilökohtaisesti koen, että tää on hyvää datanhallintaa. Mulle hyvää datanhallintaa on se, no se riippuu aina projekteista, mutta käytännössä se, että jos jossain pitää olla jotain dataa, niin se data on siellä. (...) jos se on yhteismitallista se data, niin sitä on hyvin hallittu."*

Useampi haastateltava totesi, että hyvän datanhallinnan ja erityisesti datan laadun merkitys korostuu viimeistään silloin, kun siirrytään tekoälyratkaisujen kokeiluvaiheesta tuotantokelpoisen ratkaisun kehittämiseen. *"Tavallaan on kiva rakentaa poc:ja, jotka perustuu pieniin oppimissetteihin, mutta sitten kun sen pitäisi olla operatiivista isossa skaalassa, niin siinä vaiheessa alkaa perusdatan laatu ratkaisee aika paljon."* Olematon datanhallinta ja huono datan laatu johtaa yleensä tekoälyhankkeen venymiseen tai kaatumiseen. *"Yleensähän se menee niin, että jos se on huonolla tolalla se datanhallinta lähtökohtaisesti, niin niitä ongelmia ratkotaan tekoälyprojektin tai -pilotin puitteissa, että sitten siellä värkätään data semmoiseen kuntoon, että sen kanssa jotain voidaan tehdä, mut aika usein kyllä käy niinkin, että sitten se pilotti tai projekti, kun ajatellaan, että kehitetään joku tekoälyratkaisu, niin se jää ihan alkumetreille sen takia, että todetaan aika nopeasti, että ei meillä ole tämmöistä dataa tai me ei tiedetä, onko meillä sellaista dataa tai että se on niin huonoa, että ei me voida käyttää sitä. Jonkun verran on tutkimuksiakin, että se on yksi yleisimpiä syitä siihen, että tekoälyn tai jonkun algoritmin hyödyntäminen ei onnistu, kun ei ole tai ei tiedetä, onko tarpeellista dataa saatavilla."* Huono datanlaatu johtaa siihen, että

tekoälyasiantuntijat joutuvat käyttämään valtaosan työajastaan paikalliseen datan laadun parantamiseen varsinaisen liiketoimintahyötyä tuovan toiminnan sijaan. *"...joudut käyttää ensin 90 % ajasta siihen, että yrität saada datan johonkin semmoiseen kuosiin, että sitä voi ylipäätään käyttää mallin opetukseen."*

Jotta datan laatua ei paranneta vain hankekohtaisesti tiettyä yhtä tarkoitusta varten, vaan pyritään kestävään datan laadun parantamiseen, minkä seurauksena myös uusi data on alusta asti vaatimustenmukaista ja siten suoraan laajemminkin hyödynnettävissä, datanhallinnan käytäntöjen tulee olla olemassa olevia ja riittävällä maturiteettitasolla. Jos dataa ei ymmärretä, on hankala ymmärtää tekoälyn datamuutoksista johtuvaa poikkeavaa toimintaa. *"Sinällään se merkitys menee eksponentiaalisesti, koska mun mielestä suurin ongelma noissa tekoälyhankkeissa (...) on juuri se, että ajatellaan vaan, että raavitaan jostain se data kokoon ja kokeillaan. On niin helppo tehdä erilaisia tekoälyjuttuja ja raapia sitä dataa vaan jostain ja näyttää cooleja juttuja, mutta sitten kun haluttaisiin oikeasti tehdä, niin sun pitää tuntea se data, sun pitää tietää, sen pitää olla laadukasta nyt, mutta sen pitää olla laadukasta vielä kuukauden ja vuodenkin päästä. (...) ollaan opetettu tekoäly tekemään jotain päätöksiä ja sitten tapahtuukin jotain tämmöistä uutta ja mullistavaa, niin halutaanko me, että se vaikuttaa siihen myös tulevaisuudessa vai pitäisikö meidän tehdä jotain poikkeamia, että sen datanhallinnan pitää olla tosi kovalla tasolla, jos me halutaan, että se pyörii siellä taustalla, koska kone ei osaa ymmärtää, että hei tässä onkin kyse jostain erikoisesta poikkeuksesta."*

Organisaatioista ne tulevat saamaan kilpailuetua, jotka ymmärtävät hyvän datanhallinnan vaikutuksen tekoälykehityksessä. *"Mun ehkä isoin havainto tuosta on se, että kuka tahansa pystyy rakentamaan hauskoja pilotteja tai poc:ja, mutta siinä vaiheessa, kun mennään tuotantokäyttöön, niin usein hommat kusee sitä varten, että data on huonoa eikä ole käytäntöjä korjata sitä dataa eli huono datanhallinta estää hyvien keinoäly- tai tekoälysovellusten laajemman tuotantokäyttöön viemisen."* Jos tekoäly päästetään tuotannossa valloilleen ilman ihmislähtöistä kontrollia ja jo rakennettuja datanhallinnan hyviä käytäntöjä, riskit kasvavat ja pahimmillaan menetetään mahdollisuus puuttua enää tekoälyn toimintaan. *"...tämä toinen datan hyödyntämisen aste elikkä tekoäly iteroi sitä toimintaansa koko ajan ja jos siellä lähdetään vähän samalla tyylillä tekemään, että korjataan sitten, jos jotain tulee vastaan, niin juuri sitten, kun se on oikeasti skaalattu vaikkapa sadoille eri asiakkaille vaikkapa yksi sovellus, niin siellä ei kukaan enää pysty, et pysty rekrytoimaan semmoista määrää porukkaa, joka huolehtis sen toiminnasta ellei siellä taustalla ole kaikki systeemit jo toiminnassa oletetulla tavalla."* Jos tekoälystä haluaa organisaatiolleen kilpailuvalttia, on siis väistämättä keskityttävä datanhallintaan. *"Organisaatiossa olevat ihmiset, jotka ovat alkaneet ymmärtää tai ovat aina ymmärtäneet,*

että datan käytön tehokkuus, dataan liittyvä arvontuotanto lähtee hyvästä datanhallinnasta, niin heillähän tämä tekoälyn tuleminen on hyvä, koska tämä on viimeistään asia, joka vaatii datanhallinnan periaatteet ja perusteet kuntoon."

Ihmisen on oltava edelleen päättävässä roolissa, vaikka näennäisesti päätöksiä tekee myös tekoäly. Ihmisen pitää pysyä kartalla siitä, mihin dataan tekoälyn toiminta perustuu. Tekoäly peilaa ulos sitä, millaista dataa siihen syötetään ja kertoo siten osaltaan ihmisen toiminnan ja datanhallinnan tasosta. *"Sen takia me tarvitaan sitä datanhallintaa, jota tekee ne ihmiset, mahdollisesti koneet ihmisten päättäminä. Mutta silti ne ihmiset, siksi."* Erityisesti tekoälykehityksessä on huolehdittava siitä, ettei tieto tekoälyn toiminnasta karkaa organisaatiosta henkilöstövaihdosten yhteydessä. *"Eli ettei päästetä siihen tilanteeseen sitä kehittämistä ja operointia, sen tekoälykyvykkyyden operointia tyyliin, jos on vaikka konsultti tai ulkoistettua porukkaa siinä kehityksessä ja operoinnissa mukana, niin sitten kun henkilöstö vaihtuu, niin yhtäkkiä katoaa kaikki näkyvyys siihen, että mitä tää on syöny, mistä se tieto tulee, mitä sille tapahtuu ja minne se menee, minkälaisia päätöksiä sen perusteella tehdään."* Operatiivisten roolien lisäksi organisaatiossa täytyy määritellä liiketoimintavastuulliset roolit. *"Soisi, että näissä organisaatioissa oikeasti kiinnitetään huomiota siihen, että sitä ei vaan voi nopean hyödyn tavoittelussa ostaa ulkopuolelta ja olla ymmärtämättä ollenkaan, että mikä vastuu tekoälyn käyttöön liittyy. Että sitä vastuuta ei voi ulkoistaa, se täytyy omistaa siellä organisaatiossa, joka tekoälyä hyödyntää."* Organisaation kulttuuria on kasvatettava kohti dataohjautuvuutta, mikä itsessään tukee tekoälyn vastuullista hyödyntämistä. *"Ja kaikki päätöksenteko, että mitä dataa, miten sitä manipuloidaan, käytetään ja mihin niitä lopputuotoksia hyödynnetään, niin se pitää olla sellaista bread and butter, jokapäiväistä päätöksentekokoneistoa ja sitä arviointia, että mikä on sallitun rajoissa."*

Tekoälyä hyödyntävän organisaation pitää ehdottomasti ymmärtää, mitä dataa tekoälylle syötetään ja minkä laatuisena. Monella organisaatiolla on kuitenkin parannettavaa jo datansa ymmärtämisessä. *"Jollain tavalla tänä päivänäkin vielä tosi harva organisaatio tuntee oman datansa hyvin."* Kun data ymmärretään, tekoälyä varten data yleensä kerätään keskitetysti datavarastoon. *"Tekoäly hyötyy todella paljon siitä, että saadaan yhteen paikkaan kerättyä eri tyyppisiä tietoja, että monesti tekoäly tavalla tai toisella liittyy siihen, että datasta etsitään jotain säännönmukaisuuksia, jotka sitten on kiinnostavia esimerkiksi oppimisen näkökulmasta tai ennustamisen näkökulmasta. On niin monta eri tapaa, millä tekoäly hyötyy siitä, että sillä on mahdollisimman rikas datasetti, minkä se tekoäly ottaa käsiteltäväkseen."* Kun tiedetään keskitetysti, mitä dataa tekoälylle syötetään, pystytään ymmärtämään paremmin tekoälyn tuloksia ja tuotoksia. *"...tekoäly onkin sitten sellainen, että se matematiikka siellä taustalla joissakin tapauksissa on hyvin paljon*

monimutkaisempaa ja sitten se ei ole kiveen kirjattujen sääntöjen mukaan välttämättä ollenkaan ne tulokset, vaan ne elää sisäänmenevän datan mukana."

Organisaation datanhallinnan maturiteetti ei voi olla olematonta, vaan maturiteetilta vaaditaan jonkinlaista minimitasoa, jolla voidaan turvallisesti ja rohkeasti lähteä tuotantokelpoisen tekoälyratkaisun kehittämiseen. Datanhallinnan maturiteettitaso määrittää, millaisiin hankkeisiin voidaan datan suhteen lähteä. *"...on tosi tärkeetä, että datanhallinnan ymmärrys ja maturiteetti on hyvä ja että roolit ja vastuut on selvänä ja että myös se dokumentaatio tiedosta ja tiedonkäsittelystä tekoälyn keinoin, että maturiteetti on korkea, koska muuten mä nään, että on aivan turhia ja suhteettomia uhkia."*

Jos tekoälyn taustalla on huonosti hallittua dataa, tekoälykehitykseen laitetut resurssit heitetään hukkaan ja tuloksena on liiketoimintaa jopa rampauttavaa tekoääliötoimintaa. *"Se tekoäly valehtelee niille ja se antaa niille puolitotuuksia. Luottamus tekoälyyn menee, jos se data siellä taustalla ei ole juuri sitä, mitä tekoäly tarvitsee ja niin ku me tiedetään, niin tekoäly ei osaa kuten ihminen arvioida ainakaan alussa, että onko tämä data sitä oikeata vai ei."*

Hyvästä datanhallinnasta tekoälykehitykselle koituvia hyötyjä voidaan jakaa lyhyen- ja pitkän aikavälin hyötyihin. Lyhyellä aikajanalla datan löydettävyys ja laatu paranevat ja jos datanhallinta on kunnossa tekoälykehitykseen lähdettäessä, kehitys ei pysähdy datan saatavuus- ja laatuongelmiin. *"Tokihan jos se on valmiiksi jo hyvällä tolalla, niin sitten päästään näissä tekoälyhankkeissakin nopeammin näyttämään. Stuck in experimentation -vaiheeseen ei jäädä, koska valmiiksi käytetään sitä dataa, joka on oikeaa ja hyvää ja joka päivittyy jatkuvasti."* Hyvä datan laatu näkyy nopeampana arvontuotantona organisaatiolle. *"Ainakin jos sulla on hyvää dataa, niin pääsisit heti tai pääset nopeammin rakentamaan jo ja testaa malleja ja tuottamaan jotakin näkymiä, että tässä olisi tämmöinen malli ja näin se toimisi."*

Lisäksi koordinoidulla datan hallinnoinnilla voidaan karsia päällekäistä kehitystyötä ja toisaalta auttaa identifioimaan parhaiten liiketoimintahyötyä tuovia hankkeita. *"Eli mä näen lyhyellä tähtäimellä sen, että ylipäätänsä päällekkäinen kehitystyö lähtee pois ja toinen, että siihen dataan puhtaasti kohdistuvat investoinnit ja muut, niin ne pystytään priorisoimaan järkevästi ja ehkä sitä kautta myös löytämään joitakin semmoisia matalalla olevia hedelmiä, mitä muuten ei osattaisi eikä nähtäis."* Kattavalla datan ja tekoälyn dokumentoinnilla mahdollistetaan kehitysprosessin helpompi toistettavuus ja tekoälyn uudelleenhyödyntäminen. *"... metatieto, niin sekin on tietysti siinä sitten tärkeää, että*

se buustaa kehittämisen prosessia, että koska kehittämiseen liittyy aina hirveästi tiedonsiirtoa, niin se on tehokkaampaa se kehittäminen, jos on hyvät dokumentaatiot ja metatiedot."

Pitkällä aikavälillä hyvä datanhallinta tuo liiketoiminnalle hyötyä arvon ja arvonnousun kautta. *" (Pitkän aikavälin hyödyt) tulee sitten varmaan sieltä kilpailukyvyn kautta. Että toisaalta sen datan ja datan hyödyntämisen kautta yritys voi saavuttaa erilaista kilpailukykyä markkinalla, että joko pystyy kehittää parempia uusia tuotteita tai toimii tehokkaammin ja enemmän automatisoidusti."* Hyvän datanhallinnan kautta datavarannoista on mahdollista ottaa yhtenä pääomana kaikki irti. *"...datan kehitys priorisoidaan, sitten meidän datavarannot oikeasti tulee saataville. Se liittyy siihen kehitykseen, se liittyy moneen muuhunkin asiaan. Me pystytään sitä dataa esimerkiksi katalogisoimaan, me pystytään tuomaan sinne laatu, pystytään tuomaan tämmöisiä asioita ja sitä myöten datan käyttö alkaa organisaatiossa elää ja siitä aletaan pystyä tekemään niitä hyötyjä."*

Olemassa olevat hyvät datanhallinnan käytännöt varmistavat sen, että tekoälyratkaisu on pitkäikäinen ja että myös tekoälysovelluksen tuottamaa dataa hallitaan asian vaatimalla vakavuudella. *"Datan laatu pitää pysyä hyvänä. Myös sen tekoälysovelluksen tuottaman datan oikeellisuutta, järkevyyttä on hyvä arvioida pitkällä aikavälillä, että on jotain tsekkejä, että se tekoälysovellus, joka itsessäänkin todennäköisesti tuottaa jonkinlaista käsiteltyä dataa, niin että sekin on tarkoituksenmukaista. Jos datan laatu pysyy hyvänä, niin tekoälysovellus pysyy relevanttina pidempään."* Hyvä datan laatu mahdollistaa myös uusien kyvykkyyksien rakentamisen ja tekoälyn hyödyntämismahdollisuuksien laajentamisen. *"Paljon parempia tuloksia saadaan sillä, että on se perusta kunnossa, jonka päälle tekoäly pystyy tuomaan uusia kyvykkyyksiä ja hyödyntämään sitä paremmin ja nivomaan myös osaksi, on se sitten yrityksen asiakastietoa tai yrityksen ydintuotetietoa. Ja tuo, et on tracebility ja läpinäkyvyys siitä, että mitä tekoäly tuottaa ja hyödyntää, että se on aina tavallaan jäljitettävissä siihen, vaikka yrityksen tuotetietoon tai asiakastietoon."*

Lisäksi pitkällä tähtäimellä hyvä datanhallinta voi avata myös kollektiivisia liiketoimintamahdollisuuksia organisaation ulkopuolisten tahojen kanssa. *"Tunnistetaan tärkeimmät prosessit, parannetaan niiden laatu ja pitkällä tähtäimellä päästään siihen, että me pystytään hyödyntää keinoälyä yhä laajemmin, me pystytään esimerkiksi yhdistelemään ja jakamaan dataa muiden kuin oman organisaation kanssa. Et kyllä siinä on sekä lyhyellä että pitkällä tähtäimellä ihan valtavasti hyötyä."*

4.3 Tekoälykohtaisen datanhallinnan maturiteetin arviointi

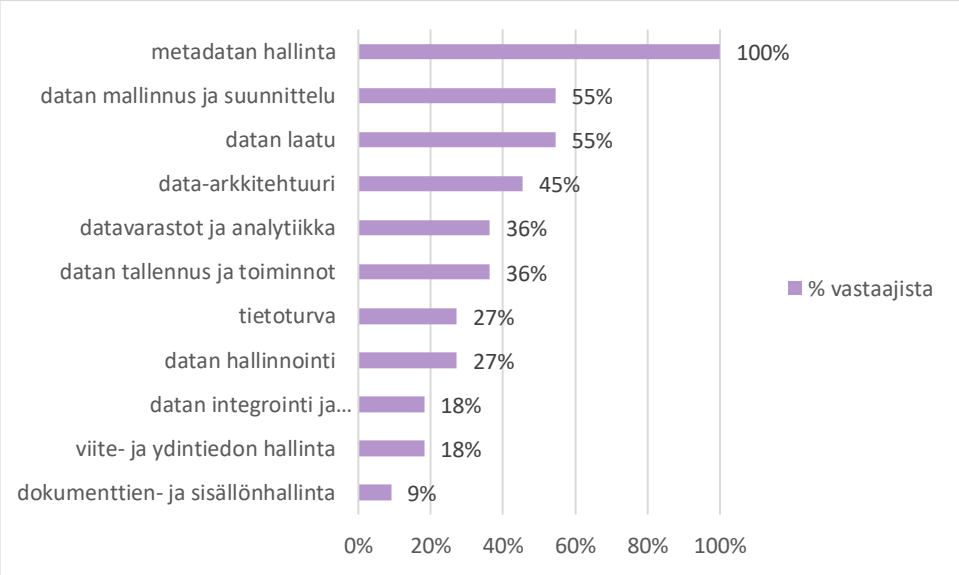
Kun arvioidaan organisaation datanhallinnan maturiteettia tekoälyn näkökulmasta, arviointi kohdistetaan niille liiketoiminnan prosessialueille, joilla tekoälykehitystä tehdään ja joista dataa tarvitaan tekoälykehitykseen. Haastateltavilta pyydettiinkin arviota vaaditusta datanhallinnan maturiteetista riippumatta alueesta, jossa tekoälykehitystä tehdään, olkoon se esimerkiksi yksi liiketoimintaprosessi tai koko organisaatio. Lisäksi eri datanhallinnan osa-alueet ja niiden kehittämisen tärkeys korostuvat tekoälykehityksen eri vaiheita ajatellen. Täytyy kuitenkin huomioida, että jokainen datanhallinnan osa-alue koostuu pienemmistä osista, esimerkiksi tietynlaisesta dokumentaatiosta, jolloin näiden välisen kehittämisen tärkeys voi vaihdella osa-alueen sisälläkin tekoälykehityksestä riippuen.

Datanhallinnan osa-alueiden vaadittu maturiteetti tekoälykehityksen ideointivaiheessa

Tekoälykehitykseen lähdettäessä haastateltavat kokivat tärkeäksi erityisesti metadatan hallinnan, datan mallinnuksen ja suunnittelun, datan laadun sekä data-arkkitehtuurin kehittämisen ennalta riittävälle maturiteettitasolle (kuvio 3). Kaikki vastaajat mainitsivat metadatan hallinnan kehittämisen olevan tärkeää ideointivaihetta ajatellen ja tämä havainto vahvistettiin myös ideointityöpajassa. Myös havainto data-arkkitehtuurin sekä datan mallinnuksen ja suunnittelun tärkeydestä validoitiin. Sen sijaan datan laadun hallinta herätti ideointityöpajassa keskustelua. Osallistujat korostivat datan laadun osalta enemmän ymmärrystä datan laadun tilasta tekoälykehityksen alkuvaiheissa varsinaisen datan laatuun kohdistuvan maturiteettivaatimuksen sijaan.

Edellisten lisäksi nostettiin esiin myös datan hallinnoinnin, tietoturvan sekä datan integrointi- ja yhteentoimivuus -osa-alueiden tärkeyttä ideointivaiheessa. Näistä datan hallinnointi sai työpajassa enemmistön kannatuksen ja lisäksi myös teoriatausta korostaa tämän osa-alueen kriittisyyttä, joten se on nostettu yhdeksi tekoälykehityksen ideointivaiheen tärkeimmistä datanhallinnan osa-alueista. On otettava huomioon, että haastatteluissa esitetyssä DAMA:n rataspyörässä datan hallinnointi sijoittuu keskelle, jolloin tämä osa-alue oli muita helpompi jättää huomioimatta. Osa haastateltavista kommentoi datan

hallinnoinnin mahdollista tärkeyttä tekoälykehityksessä vasta, kun asiasta muistutettiin haastattelun yhteydessä.



Kuvio 3. Tärkeimmät datanhallinnan osa-alueet tekoälykehityksen ideointivaiheessa

Kun organisaatio lähtee ideoimaan tekoälyn hyödyntämistä, sen taustalla on yleensä tavoite ratkaista jokin ongelma tai kartoittaa uusia liiketoimintamahdollisuuksia. Sen lisäksi, että mietitään taloudellisten resurssien kautta mahdollisuutta lähteä kehittämään tekoälyä, myös data on huomioitava tarvittavana resurssina tekoälykehityksessä. *”Ideointivaiheessa on siis tärkeätä se, että jos lähdetään kartoittamaan vaikka jonkinlaisia pullonkauloja prosesseissa, vaikka joihin voisi tekoälyllä pureutua, niin täytyy ymmärtää se, että minkälaista dataa ne erilaiset sovellutukset saattaisivat tarvita ja sitten täytyy määritellä se, että onko meillä sitä dataa ylipäätään ja päästäänkö me käsiksi siihen dataan mitenkään.”*

Dataa on ylipäätään mahdoton hallita, saati ideoida datan pohjalta, jos datasta ei ole kuvauksia. *”Data-arkkitehtuuri on ainakin niiltä osin pakko olla erittäin hyvä, että se on oikeasti pakko olla kuvattu, koska muuten meillä ei ole datan elinkaarenhallintaa, että senkin panos on tietyssä mielessä iso.”* Jos datasta löytyy ajantasaisia tietomallikuvauksia, datarakenteet ja siten dataelementtien yhteydet toisiinsa hahmottuvat nopeammin ja voidaan keskittyä varsinaiseen ideointiin. *”Että jos sinulla ei ole mitään käsitystä, että mistä se data koostuu, niin se on aika hankala ideoida. Sä voit vaan arvailla. Et kyl mä nostaisin ehkä tuon data-arkkitehtuurin ja datan mallinnuksen aika tärkeäksi osaksi tuota ideoinnin onnistumista.”* Ideoinnissa voidaan hyödyntää ylätason eli konseptuaalisen tason tietomallinnusta, jolla voidaan hahmotella myös uusia tarvittavia, vielä puuttuvia dataelementtejä. Mallinnuksen pohjalla on kuitenkin hyvä olla olemassa olevat kuvaukset

nykytilasta. *”Data-arkkitehtuuri varmaankin myös eli mitä dataa jo on ja mitä dataa puuttuu, jotta siitä tekoälyn ajatellusta höydystä päästäis nauttimaan. Ideointivaiheessa just varmaan ylipäänsä se, että ymmärtää sen nykytilan.”*

Ennakoivalla metadatan hallinnalla saadaan tarkempaa tietoa siitä datasta, jota lopulta käytettäisiin tekoälykehityksessä. *”Todennäköisesti se kaikkein tärkein asia on metadatan hallinta, koska se tarkoittaa mulle nykyaikana aika paljon sitä data katalogia. (...) Silloin me tiedetään, että mitä tietoa meillä oikeasti on olemassa, niin silloin se metadatan hallinta on ehkä kaikkein tärkein tossa ideointi- ja oikeastaan myöskin AI-projektin määritysvaiheessa. Meidän pitää tietää, mitä dataa meillä on, ihan oikeasti, että mitä se on ja sen jälkeen sitten vasta tulee kaikki noi muut.”* Koska data on edellytys tekoälylle, tieto datasta on ennakkoedellytys tuotantokelpoisen tekoälyratkaisun ideoinnille ja konseptoinnille. *”...sulla pitää olla tietyt standardit ja laatusäännöt määritelty ja data kuvattu, jolloin sun on paljon helpompi tehdä sitä ideointia. Mä näkisin, että se (maturiteetin) kolmostaso voisi olla sitten se, että tekoälyn ideointi on mahdollista tehdä.”*

Jos kuvaukset tiedosta puuttuvat ja ideointia joudutaan tekemään tietokannoista käsin, liiketoiminta on vahvasti riippuvainen IT-puolen osaajista eikä ideointi välttämättä silloin palvele liiketoimintaa sillä tasolla kuin olisi mahdollista, jos kuvaukset datasta olisivat liiketoiminnalle ymmärrettävässä muodossa. *”No ideoinnissa varmaan, jos sinulla olisi hyviä kuvauksia datasta olemassa, niin sun ei välttämättä tarvis päästä vielä kovin syvälle sinne kantoihin niitä tutkimaan. (...) Sitten ideointia voi tehdä ehkä suurempi joukko ihmisiä, jos sinulla on hyvät kuvaukset siitä datasta olemassa. Versus että jos mitään ei oikein siitä datasta ymmärretä ja sä tarviit jo siihen ekspertin tai teknisen ihmisen, joka pystyy alkaa selvittää, että mitä ehkä me voitaisiin tehdä, mitäs tämä meidän data on.”* Tietokantalähtöinen ideointi ei myöskään anna kokonaiskuvaa kaikista mahdollisista tekoälyratkaisun hyötymistä datoista. Tieto datasta on ensimmäinen askel siihen, että organisaatiossa jo makaava data on hyödynnettävissä myös muualla liiketoiminnassa. Metadatakuvaukset antavat vihjeitä siitä, että dataa hallitaan, jolloin data on sitä kautta myös todennäköisemmin luotettavaa. *”Siinä ideointivaiheessa on tärkeää tietää, että mitkä on käytettävissä, mitkä on luotettavia.”*

Jotta aiemmin mainitut datanhallinnan osa-alueet on kehitetty sellaiselle tasolle, että ne oikeasti palvelevat organisaation liiketoimintaa, tarvitaan lisäksi tätä kehittämistä edistävää ja valvovaa strukturoitua toimintaa. *”Eli siis meillä pitää se keskimmäinen pallukka olla olemassa ihan alusta asti jollakin kypsyysasteella ja myöskin laajuusasteella, että sen ei tarvitse olla koko organisaation tasosta, mutta sen pitää olla olemassa jonkun sorttista niin kuin datan hallintaa, data governance -hallintaa.*” Lisäksi identointityöpajassa korostettiin,

että alusta asti tarvitaan datan hallinnoinnin tuomat raamit datan hallinnalle tekoälykehityksessä, jotta suunnitelmat perustuvat faktoihin ja ymmärretään se, onko data ylipäänsä hyödynnettävissä myös oikeutuksen ja luvallisuuden kannalta. Esimerkiksi, jos hankkeessa tullaan käyttämään henkilötietoja, tietoturvan ja siten tietosuojan maturiteettitaso tulee myös olla kohdillaan. *”Tietysti tietosuoja on vähän sellainen, että jos alkuvaiheessa jo tiedetään, että tässä tullaan käyttämään jotain henkilötietoja, niin se arviointi tietysti täytyy tehdä heti alkuvaiheessa.”*

Lisäksi datan tallennus- ja toiminnot -osa-alueen osalta on syytä nostaa esiin se, että monet tekoälyn käyttötapaukset vaativat kerättyä dataa pitkältä aikaväliltä, jotta saadaan tekoälystä toivottua hyötyä. Ideointivaiheessa tulee siis kerätä tarkoin vaatimukset datalle. *”...keinoälys varsinkin monet laitteet on sellaista, jos ne vaikka hajoo, ne saattaa hajota kerran kymmeneen vuoteen tai kolmeen vuoteen. Tällaisissa dataa täytyy olla mitattu tosi pitkältä ajalta. Ja se on yksi kanssa siinä mielessä iso ongelma, että jos organisaatio herää siihen, että haluaa nyt ruveta kehittää jotain vaikka predictive maintenance:a ja sitten ne laittaa sinne mittareita, sensoreita mittaa dataa, niin niillä voi olla vaikka puolelta vuodelta dataa ja se voi olla aika rajoittunut, mitä sille voi sit edes tehdä.”*

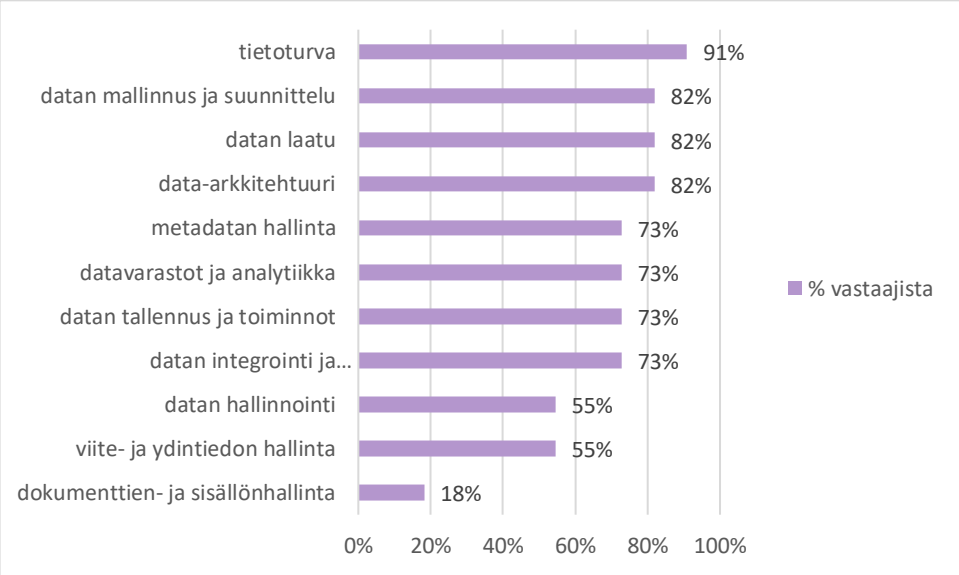
Vastaajista 73 % antoi arvionsa metadatan hallinnan vaaditulle maturiteettitasolle ideointivaiheeseen lähdettäessä. Näiden vastausten keskiarvo oli kolme eli ennakoiva maturiteettitaso. *”Kun se on sillä alueella missä se data, joka siihen AI:hin liittyy, niin silloin se riittää sillä alueella, että on kolme.”* Lisäksi haastatteluissa ja työpajassa vastaajat arvioivat, että myös datan hallinnoinnin pitäisi olla ennakoivaa ja määriteltyä eli maturiteettitasolla kolme tässä vaiheessa. *”Kyllä pitäisi (data hallinnoinnin maturiteettitason olla alkuvaiheessa) kolmonen suurinpiirtein varmaan, että se on oikeasti strukturoitua ja systemaattista ja toimivaa, ettei se ole ad hoc:ia, että tehdään jos muistetaan tai varsinkaan, että se puuttuu kokonaan, mutta että varmaan vähintään siellä kolmostasolla.”*

Kun datasta on kerätty tarpeeksi informaatiota ideointia varten ja ideointia kohdistetaan tiettyyn dataan, datan laadun taso olisi hyvä ymmärtää. *”Että meillä on ymmärrys siitä laadusta, on tärkeä siinä ideointivaiheessa. Tavallaan se, mitä kaikkea suunnitellaan, niin ettei se ole liian kaukana sitten toteutuskelpoisesta.”*

Datanhallinnan osa-alueiden vaadittu maturiteetti AI-projektin määritysvaiheessa

AI-projektin määritysvaiheessa kaikkien datanhallinnan osa-alueiden olisi syytä olla vastaajien mielestä maturiteettitasolla kolme dokumenttien- ja sisällönhallintaa lukuun ottamatta (kuvio 4, 44). Kaiken tekoälykehityksen piiriin kuuluvan datanhallinnan tulisi siis

olla ennakoivaa ja määriteltyä. Jo ideointivaiheessa tärkeiksi koettujen datan mallinnuksen ja suunnittelun, data-arkkitehtuurin, metadatan hallinnan ja datan hallinnoinnin kehittämisen tärkeyttä AI-projektin määrittelyvaiheessa korosti aiempaa suurempi vastaajien joukko. Uusina korostettuina datanhallinnan osa-alueina nousivat erityisesti tietoturva ja datan laatu. Vastaajista 91 % korosti tietoturvan ja sen sisältämän tietosuojan huomioimista tässä vaiheessa. Datan laadun osalta riittävää maturiteettitasoa korosti 82 % vastaajista.



Kuvio 4. Tärkeimmät datanhallinnan osa-alueet AI-projektin määrittelyvaiheessa

AI-projektin määritysvaiheessa tulee ymmärtää, sisältääkö hanke arkaluonteisen datan, kuten henkilötietojen hyödyntämistä. *"...AI-projektin määritysvaiheessa pitäisi olla tunnistettu, että miten meillä käsitellään tätä ja miten sitä saa käyttää. Ja siinä vaiheessa, kun tiedetään, mitä projektia ollaan tekemässä, niin silloin varmaan tiedostetaan se tietoalue, domainit ja onko esimerkiksi henkilötietosuoja-scope siellä mukana vai ei ja niin edespäin, niin kyllä se tossa projektin määritysvaiheessa täytyy olla jollain tapaa tunnistettu."* Tähän kytkeytyy vahvasti myös jo ideointivaiheessa korostettu metadatanhallinta. Ilman kattavaa dokumentaatiota datasta ei voida saada ymmärrystä siitä, mitkä dataelementit, -attribuutit ja -kentät mahdollisesti sisältävät arkaluonteista dataa, jonka hyödyntämiseksi tarvitaan säännöstöjä ja toimenpiteitä. *"Toki ensin, että sinulla on se idea, että mitä yrität tehdä, mutta sitten että sä oikeasti ymmärrät sen datan, mikä sulla on käytössä."* Jos datanhallinnan toimenpiteet jättää tekemättä ja kehittämättä ennakoivalle maturiteettitasolle, varsinaisissa hankkeissa ja projekteissa datanhallinnan puutteet näkyvät aikataulujen venymisenä. *"No se (pullonkaula muodostuu) varmaan siellä alkupäässä, että sitä dataa ei ole*

kuvattu. Että sulla on vaikea vaan saada ymmärrys ensiksi siihen, että mitä tämä data tarkoittaa ja mitä nämä kentät, mitä nämä arvot täällä kentissä on, vaikka sinänsä olisikin kohtuu ymmärrettävä juttu niin ehkä siihen vaan menee ensin paljon aikaa."

Jos AI-projektissa päädytään hyödyntämään arkaluonteista dataa, tarvitaan hyvin määriteltyä pääsyn- ja oikeuksienhallintaa. *"Itse asiassa tietoturva ja tietosuoja varmaan on otettava huomioon, oli se sitten tuo AI-projektin määritysvaihe tai datan kuratointi, mutta siinä vaiheessa, kun pitäisi jotakin sille datalle tehdä, että kuka siihen saa päästä käsiksi ja millä oikeuksilla. Ehkä sen nostaisin tässä molempiin näistä vaiheista."* Tietoturvan ja tietosuojan ennakoiva maturiteettitaso tekoälykehityksen vaikutusalueilla toimii eräänlaisena tarkastuspisteenä kohti luottamuksenarvoista tekoälyratkaisua. *"...sehän pitäisi olla sitten hallussa että voidaan ampua ne projektisuunnitelmat alas jos ne esimerkiksi ei ole laillisia."* Tietoturvan ja tietosuojan korkeampi maturiteetti laskee organisaation riskitasoa. *"...jos datanhallinta on olematonta tai kuralla ja halutaan hyödyntää tekoälyä, niin näen, että varsinkin tietosuojanäkökulmasta ja ehkä muutenkin tavallaan että mitä hyötyä tekoäly ylipäänsä voi tuottaa, niin jos datanhallinta ei ole hyvällä tolalla, niin siitä tekoälystä voi olla enemmän haittaa eli tietosuojamielessä siitä voi olla todellisia uhkia rekisteröityjen oikeuksille ja organisaatiolle itselleen siitä vinkkelistä, että onko oikeasti ymmärretty ja dokumentoitu, että mihin käyttötarkoituksiin ja millä keinoin henkilötietoa käsitellään ja onko pystytty informoimaan asiakkaita, rekisteröityjä asianmukaisella tavalla siitä, että miten ja mihin sun tietoas tekoälyn keinoin käsitellään."*

Datan laatu on yksi merkittävimmistä asioista, mikä tulee selvittää AI-projektin määritysvaiheessa. Datan laatuvaatimukset määritetään aina kunkin tekoälykehityshankkeen mukaisesti. *"Täytyy olla hyvin määritelty se datan laatu, minkä laatuista dataa sinne tekoälysovellukseen voidaan syöttää sisään, että mun mielestä se on esimerkiksi yksi semmoinen merkittävä asia."* Tekoälykehitykseen lähdettäessä on otettava huomioon mahdollinen hankkeen rinnalla tai yhteydessä tehtävä datan laadun kehittäminen siten, että data palvelee tekoälyä kestävästi pitkällä tähtäimellä. *"Sitten tietenkin datan laatu, että pitääkö sun keskittyä jonkun tietyn tietoalueen, vaikka datan laadun putsaukseen ensin ennen kuin sä voit tehdä jonkun hyvän tekoälyhankkeen vai onko asiat niin hyvin, että tavallaan pystytään lähteä etenemään."* Jälleen kerran, jos datan laadun eli tekoälyn elintärkeän polttoaineen kehittämiseen ei keskitytä, tekoälyhankkeen riski epäonnistua ja venyä loputtomiin kasvaa. Tällöin saatetaan päätyä nopeisiin pikaratkaisuihin ja korjata dataa paikallisesti ilman, että datan laatu paranisi kestävästi pitkällä tähtäimellä. *"Toki siinä projektin määritysvaiheessa meidän olisi hyvä ymmärtää, että mikä se laatu todellisuudessa on. Koska sitten kuratoinnin määrä riippuu datan todellisesta laadusta. Eli me*

joudutaan sitten miettimään sitä, että kuinka paljon sitä kuratointia joudutaan tekemään, riippuen siitä, että mikä se laatu oikeasti on sille datalle mitä me tarvitaan."

Vastaajista 73 % piti tärkeänä, että datanhallinnan osa-alueista datan tallennus ja toiminnot, datavarastot ja analytiikka sekä datan integrointi ja yhteentoimivuus ovat kehitetty ennakoivalle maturiteettitasolle AI-projektin määritysvaiheessa. Kun tekoälyn tarvitsema data on määritelty ja ymmärretään, missä kyseistä dataa on tallennettuna ja minkä laatuisena, on mietittävä, miten dataa saadaan mahdollisesti eri järjestelmistä ja myös ulkoisista lähteistä tekoälyn käyttöön. *"Ja sitten tietenkin projektin määritysvaiheessa on tärkeä tietää datan tallennuspuoli ja se, että missä sitä on saatavilla. Eli hahmotelma siitä, että kun me lähdetään tekemään tällaista hanketta niin mistä järjestelmästä sitä tultaisiin ottamaan. Mihin sitä mahdollista tekoälyhanketta tultaisi kytkemään tai pilotoimaan."* On siis tarkasteltava tekoälyn käyttämän datan elinkaarta ja mietittävä, pitääkö dataa tallentaa keskitetysti muualle, jotta se olisi paremmin tekoälykehityksessä hyödynnettävissä. *"Datan tallennus ja toiminnot, joo. Kyllä määrittelyvaiheessa meidän pitää jo muodostaa näkemys siitä, että missä sitä on, mihin me halutaan sitä tallentaa, miten tallennetaan."* Tekoälykehityksessä hyödynnettävän datan elinkaaren suunnitteluun on panostettava, jotta ollaan tietoisia tekoälyn perustana olevan datan tilasta. Tällöin mahdolliset datamuutokset tehdään tietoisesti eikä tiedostamatta, miten tietyt aktiviteetit vaikuttavat tekoälyn toimintaan. *"Eli mitä jos jotain tapahtuu niin voidaanko me sanoa, että data ei ole näissä paikoissa muuttunut."*

Jos tekoäly tulee hyödyntämään useasta järjestelmästä tulevaa tietoa, integrointikyvykkyyden on oltava sellaisella maturiteettitasolla, että tekoälylle voidaan varmistaa oikeaaikaisen ja oikeamuotoisen datan saaminen. *"Integroinnista pitää ottaa jo selkeä suunnitelma tuossa vaiheessa."* Datojen yhteentoimivuus on oltava hyvällä tasolla. *"Miten mä saisin ne datat yhdistettyä."*

Jos tekoälykehityksessä hyödynnettävä data keskitetään, datavarastojen analyyttiset kyvykkyydet on huomioitava datavarastoa valitessa. *"Jos se on kovin vanhakantainen, perinteinen datavarasto, niin sitten kun sä laitat siihen jonkun hirveän algoritmihirviön kiinni, niin ei se kestä, ei se pysty palvelee sellasta."* Organisaation analytiikkatoiminnot tulisi olla jo pitkälle kehitettyjä. *"Keinoäly varmaan on niin hyvä kuin analytiikka, johon se pohjautuu."* Hyvä analytiikka tuo paremmin näkyville myös sen, mitä liiketoiminnallista hyötyä tekoäly voisi tuoda. *"Analyyttinen tekeminen pitää olla hanskattu tuossa vaiheessa, että sä pystyt määrittelemään sen hyödyn, mitä sille AI-projektille tulisi."*Toki, jos tekoäly käyttää esimerkiksi vain yhden järjestelmän tai yhdestä sensorista tulevaa dataa eikä

tulevaisuudessa ole suunnitelmissa laajentaa tekoälyn hyödyntämismahdollisuuksia, datavarastoratkaisut eivät ole pakollisia. *”Data voidaan ottaa datavarastosta tai sitten se voidaan olla ottamatta.”* Joka tapauksessa päätös tiedon varastoinnista tulee tehdä AI-projektin määritysvaiheessa. *”Totta kai meidän pitää AI-projektin määrittelyvaiheessa miettiä, että mistä tietovarastot, missä tieto varastoidaan ja minkälaisia analyyttisia ratkaisuja me tehdään, että täytyyhän sen olla pitkällä jo.”* Myös pitkällä tähtäimellä ja pitkän aikavälin hyötyjä tavoiteltaessa datavarastojen ja analytiikan parissa tehtävien asioiden on oltava hallittuja, jotta ei muuteta tekoälyn hyödyntämää dataa tiedostamatta ja toisaalta kyetään varmemmin ennustamaan tekoälyn toimintaa. *”Eli mitä jos jotain tapahtuu niin voidaanko me sanoa, että meillä ei ole jonkunlaista datavarantoa tai platformia, jossa ehkä jostain syystä se datasetti saattaisi muuttua, olemme analysoineet, me pystymme analysoimaan näin, jotta voimme ennustaa bias-välin.”*

Viite- ja ydintiedonhallinnan maturiteettitaso on relevantti, jos tällaista dataa halutaan ylipäätään syöttää tekoälylle. Siksi datamassojen erottelussa muutamiksi eri osa-alueiksi ei ole tekoälykehityksen näkökulmasta välttämättä merkitystä. *”Se muuttuu AI-maailmassa epärelevantiksi, koska kaikki data pitää hallita, kaikki AI:hin liittyvä data pitää olla riittävällä tasolla ja siinä mielessä ydin- ja viitetieto on samalla viivalla kaiken muun sen AI:n tarvitseman tiedon kanssa.”* Olkoon hyödynnettävä data mitä tahansa, kyseisen datan hallinnan on oltava tekoälyhankeen laajuuden mukaan riittävällä maturiteettitasolla onnistuneen tekoälyhankkeen varmistamiseksi. *”Mä luulen, että viite- ja ydintiedon hallinta, jos on joku siiloisempi tai pienemmällä scopella oleva tekoälykehitys, niin varmaan siihen riittäis vähän vähempikin, että jotain kakkosen ja kolmosen (maturiteettitason) väliltä, mut mitä isommasta kokonaisuudesta on kyse, niin sen tärkeemmäks tuokin nousee.”* Riittävä viite- ja ydintiedonhallinnan maturiteettitaso helpottaa myös varsinaisessa tekoälyllä saavutettavan liiketoimintatavoitteen määrittelyssä. *”No siis tää on sitten just se, että mitä sä haluat tehdä sillä datalla, että minkälaista dataa haluat käyttää, että periaatteessa tää tulee ehkä sieltä business case -määrittelyn kautta, koska siihen sä tarvitset kyllä ydintietoa aika paljonkin.”*

Jo ideointivaiheessa oli korostettu datanhallinnan osa-alueista data-arkkitehtuurin, datan mallinnuksen ja suunnittelun sekä datan hallinnoinnin merkitystä. AI-projektin määritysvaiheessa kaikkien näiden osa-alueiden merkitys kasvaa, kun määrittelyä tehdään tarkemmalla tasolla. Ilman arkkitehtuurikuvauksia määrittelyvaihe uhkaa venyä tai kuvauksia tehdään virhealttiisti nopealla aikataululla. *”Sitten (AI-projektin määritysvaiheessa) lähtisin varmaan jo miettii vähän sitä, että toki nyt mun pitää tarkemmin ymmärtää, että missä se on ja sitä data-arkkitehtuuria, että mistä se data löytyy.”* Olemassa olevat arkkitehtuuriku-

vaukset ovat hyvä pohja, kun määritellään tarkemmin liiketoiminnan vaatimuksia tavoitellulle tilalle. *”Mallinnus ja suunnittelu sillä tavalla, että siinä vaiheessa me käydään niitä liiketoiminnan vaatimuksia enemmän läpi.”*

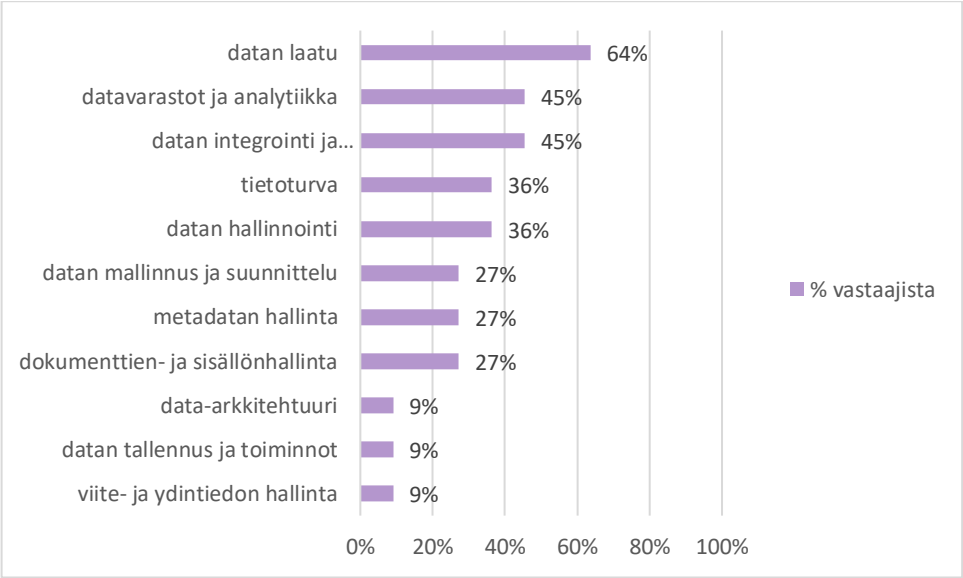
Jotta datanhallinta palvelee tekoälyä jokaisella tarvittavalla datanhallinnan osa-alueella, datan hallinnoinnin eli datanhallintaa priorisoivan ja monitoroivan toiminnan tulee myös olla määriteltyä. Lopulta ihminen vastaa myös tekoälyn ja sen taustalla olevan datan toiminnasta, joten roolit ja vastuut tulee olla selkeät. *”Sitten siinä (AI-projektin määritysvaiheessa) tulee toi datan hallinnointi. Se data governance astuu jo aika tärkeeseen rooliin, että tulee kaikki omistajuudet.”* Datan hallinnoinnin kautta saadaan nopeasti ja kattavasti tietoa siitä, onko haluttua dataa saatavilla, minkä laatuista se on ja onko se ylipäätään hyödynnettävissä. *”Jotta sä voit määritellä hyvän projektin, niin sun pitää tunnistaa, että ne tiedot, mitkä on kuvattu ja mitä oot ajatellut mitä lähtisit sitten hyödyntää siinä tekoälyhankkeessa, ni sun pitää tietää että kuinka ne prosessoidaan ne tiedot tällä hetkellä eli tullaan tähän hallintaan.”* Datan hallinnoinnin kautta datan hyödyntämiselle on sovitut prosessit, käytännöt ja säännöt, jolloin data on hyödynnettävissä laajasti koko organisaatiossa, jos ja kun esimerkiksi tekoälyä skaalataan laajemmalle. *”On kirjattu, mitä tehdään ja omistajat tietää, että tätä dataa tullaan hyödyntämään ja on selkeät prosessit, että mistä se otetaan ja mitä sille tehdään.”*

Maturiteettitasovaatimus dokumenttien ja sisällönhallinnan osalta riippuu siitä, käyttääkö tekoälyratkaisu strukturoimatonta dataa vai ei. Jos käyttää, niin on tekoälyn hyödyntämä data mitä tahansa, tämän datan hallinnan on oltava riittävällä tasolla. Varsinaista tekoälyyn liittyviä dokumentointivaatimuksia ei vielä projektin määrittelyvaiheessa niinkään ole. *”Toi dokumentaatio ei itselle näyttele hirveän kummoista roolia varsinkaan määritysvaiheessa. Toki pitää olla ymmärrys siitä, että mitä on tehty aikaisemmin ja miksi.”*

Datanhallinnan osa-alueiden vaadittu maturiteetti datan kuratointivaiheessa

Tekoälykehityksessä siirrytään AI-projektin määritysvaiheen jälkeen datan kuratointiin. Vastaajista 64 % korosti datan laadun hallinnan tärkeyttä tässä vaiheessa tekoälykehitystä (kuvio 5). Lisäksi 45 % vastaajista korosti datanhallinnan osa-alueista lisäksi

datavarastojen ja analytiikan sekä datan integrointi ja yhteentoimivuuden riittävää maturiteettitasoa tässä vaiheessa tekoälykehitystä.



Kuvio 5. Tärkeimmät datanhallinnan osa-alueet tekoälykehityksen datan kuratointivaiheessa

Datan laadun kehittämistä korostettiin sekä AI-projektin määritysvaiheeseen että datan kuratointivaiheeseen mentäessä. *”Kun dataa kuratoidaan, laatu nousee ihan äärettömän tärkeäksi.”* Datan laatu on kuitenkin syytä ottaa huomioon jo ennen datan kuratointivaihetta, jotta tekoälykehityshanke ei veny ja mahdollisesti kaadu siksi, että datan laatua analysoidaan ja korjataan vasta, kun hanke on jo pidemmällä. *”Siinä vaiheessa, kun me ruvetaan käymään sitä dataa läpi, niin sulla pitää olla hyvä ymmärrys siitä datan laadusta, että siellä se pitää olla mun mielestä proaktiivista jo, ennakoivaa.”* Datan kuratointivaiheessa siirrytään valitsemaan tekoälylle dataa, joten datan laadun merkitystä ei voi painottaa liikaa. *”Datan laatu. Mun mielestä siinä (datan kuratointivaiheessa) on tärkeintä datan laatu, datan laatu ja sitten datan laatu.”*

Erityisesti, kun dataa kerätään eri järjestelmistä datan kuratointivaiheessa, integraatiokyvykkyyden ja datan yhteentoimivuuden on oltava maturiteettitasoltaan määriteltyä. *”Datan integrointi eli onko sitten integroitu kuin laajalti. Puuttuuko sieltä joku keskeinen järjestelmä, mikä pitäisi olla mukana. Ja missä muodossa sitä on saatavilla elikkä yhteentoimivuus.”* Jos dataa kerätään keskitettyyn datavarastoon, myös tämän kyvykkyyden on oltava määritellyllä tasolla. Näiden kyvykkyyksien kehittäminen on siis aloitettava paljon aiemmin, jos tekoälylle halutaan syöttää laajemmalti kuin yhdestä järjestelmästä dataa. Datan täytyy joka tapauksessa olla yhteentoimivaa tulevia vaiheita ajatellen. *”Mehän voidaan tehdä prototyyppejä aika paljon vaikka pikku dumpeilla tai otoksilla, ettei välttämättä*

tarvitse integroida kaikkia yhteen. Yhteentoimivuutta vaaditaan siltä datalta. Totta kai sen pitää olla siinä muodossa mitä halutaan."

Kun dataa kerätään, tarvitaan viimeistään tälle prosessille säännöt siitä, kuka, miksi ja milloin dataa on luvallista kerätä. *"Itse asiassa tietoturva ja tietosuoja varmaan on otettava huomioon, oli se sitten tuo AI-projektin määritysvaihe tai tämä datan kuratointi, mutta siinä vaiheessa, kun pitäisi jotakin sille datalle tehdä, että kuka siihen saa päästä käsiksi ja millä oikeuksilla. Ehkä sen nostaisin tässä molempiin näistä vaiheista."*

Myös dokumentointiin on kiinnitettävä enemmän huomiota, kun dataa kerätään tekoälyä varten, jotta tekoälyn toiminta on alusta saakka hallittua, ennakoitavissa ja myös skaalattavissa. *"Sen pitää olla dokumentoituna, saatavilla ja ajantasaista sen tiedon, mitä meillä on siitä AI:sta."* Dokumentointi on erityisen kriittistä tekoälykehityksessä, jossa perusteet tekoälyn toiminnalle eivät ole suoraan nähtävissä tekoälystä vaan on pureuduttava tekoälyn hyödyntämään dataan ja siihen tehtyihin valintoihin. *"Tässä sä alat luomaan uusia datasettejä, niin sit sun pitäis tietenkin dokumentoida ne, mitä sä olet luonut, miten sä olet sen datan käsitellyt, mihin varastoon sä sen laitat."*

Riittävä ja ajantasainen dokumentointi on avainasemassa tekoälyn käyttämän datan hallinnoinnissa. Erityisesti työntekijöiden vaihtuessa tarvitaan dokumentaatiota datasta ja tekoälystä sekä siirtyviä datan vastuurooleja, jotta tieto tekoälyn toiminnasta ei lähde organisaatiosta työntekijöiden mukana. *"Jos sinulla on kysyttävää siitä tai datan kuratoinnissa sä olet uusi ihminen, sä saat sen datasetin ja sä mietit, että mitä tämä on. Jollakin tasolla se on varmaan kuvattu, mutta silti, onko ne kuvaukset miten ajantasalla. Sä haluat jonkun ihmisen kanssa kuitenkin jutella, että mitäs tässä näkyy ja miten nämä tänne päivittyy, niin voisihan se olla hyvä, että siinä vaiheessa olisi jotain kontaktihenkilöitä, jotka saa kiinni. Samaten, kun mennään eteenpäin tuossa, että joku ymmärtää taas siitä mallista, niin sitten sinulla olisi semmoinen haju, että kuvaukset, että täältä nämä ihmiset löytyy, jotka osaa kertoa lisää. Ja juurikin sitten varmistuis se, että jos joku lähtee firmasta niin ne vastuut siirtyisi, mikään ei putoa. Just, että mitä pidemmälle mennään sitä tuotantoa kohti, niin tärkeämpiä."*

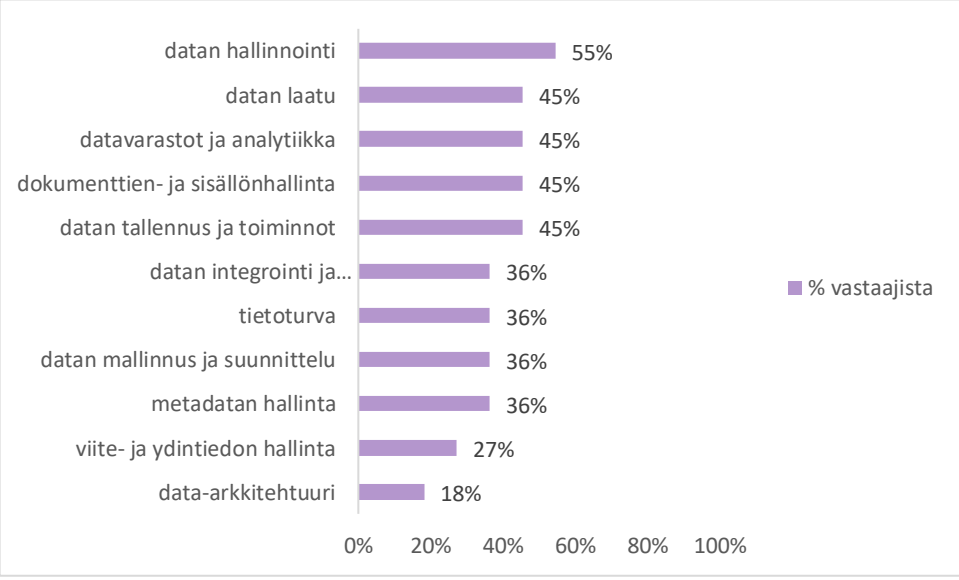
Riippuen tekoälyratkaisun laajuudesta, datan hallinnointi voidaan skaalata sen mukaan. Tärkeintä on, että datan hallinnointi ohjaa ja tukee liiketoiminta-arvoa tuottavaa datalähtöistä toimintaa. Minkäänlaisia hallintorakenteita ei kannata luoda ilman sen liiketoiminnalle tuomaa strategista arvoa. *"Että governance on sillä tasolla, että tiedetään mitä pitäisi tehdä aina, on ehkä ensimmäisiä omistajia ja strategisella tasolla ja reaktiivista toimintaa ja ehkä sillä jo päästään liikkeelle, mutta se on tavallaan oma polkunsa sitten*

sen governancen kehittäminen. En näkisi, että se kannattaa olla liikaa esteenä tekoälyratkaisulle."

Jos datan hallinnointia ei ole olemassa, mahdollisuudet varmistaa tekoälylle kattavaa ja hyvälaatuista dataa ovat pienet ja siten myös tekoälyn potentiaali jätetään käyttämättä ja sen skaalaaminen mahdollistamatta. Ilman ohjausta riskit tekoälykehityksen epäonnistumiselle kasvavat. *"Jos ei ole sitä governance:a olemassa, datan hallinnointia, niin se korjaaminen on tosi paljon vaikeampaa. Niin siinä vaiheessa, jos governance on kunnossa, se onnistuu tosi paljon tehokkaammin, että sitten jos sitä ei ole, niin se on hirveän hankalaa. Silloin siitä tulee sellaista, että sen sijaan, että tiedettäisiin kenen pakeille mennä, niin se menee siihen, että metsästetään niitä vastuita ja kuka jotain tekee, että siitä tulee sekavaa."*

Datanhallinnan osa-alueiden vaadittu maturiteetti prototyypin luontivaiheessa

Vastaajista 55 % korosti datan hallinnoinnin ennakoivan maturiteettitason tärkeyttä prototyypin rakentamisvaiheessa (kuvio 6, 51). Lisäksi 45 % korosti sekä datan laadun, datavarastojen ja analytiikan, dokumenttien- ja sisällönhallinnan sekä datan tallennuksen ja toimintojen riittävän maturiteetin tärkeyttä tässä vaiheessa. Yleisesti ottaen vastaajien mielestä kaikkien tekoälykehitykseen kytköksissä olevien datanhallinnan osa-alueiden tulisi olla kehitetty maturiteettitasolle kolme, kun siirrytään rakentamaan tekoälyratkaisun prototyyppiä. Datanhallinnan aktiviteettien on siis oltava määriteltyjä ja ennakoivia.



Kuvio 6. Tärkeimmät datanhallinnan osa-alueet tekoälykehityksen prototyyppivaiheessa

Tekoälyratkaisun prototyyppiä rakennettaessa datan hallinnointi muotoutuu tekoälyn hallinnoinniksi, koska tekoälyn toiminta perustuu vahvasti sen hyödyntämään dataan. Tekoälyä valvotaan monitoroimalla siihen perustuvaa dataa ja tekoälyä hallinnoidaan määrittämällä rooleja ja vastuita sekä sen perustalla olevalle datalle että tekoälyn tuottamille tuotoksille ja tuloksille. *”Datan hallinnointi on tuossa vaiheessa tosi tärkeätä, koska pitää kirjata governance-malliin, tehdä AI-governancea, että miten sitä valvotaan, kuka sen omistaa, kuka sen datan omistaa, kuka ne lopputulokset omistaa. Kone ei vastaa tekemistään päätöksistä, jonkun muun pitää siitä vastata.”*

Rakenteet datan hallinnoinnille ovat kriittisiä, kun rakennetaan tekoälyratkaisun prototyyppiä, jotta voidaan varmistaa, että tekoälyn toiminta pohjautuu riittävän kattavaan ja hyvälaatuiseen dataan ja että tekoälyratkaisu tukee liiketoiminnan sille asettamia tavoitteita. *”Siinä vaiheessa datan hallinnointi on siinä mielessä tärkeää, että sinulla on oikeat stakeholderit ja prosessi-ihmiset mukana sen validoimisessa, että onko se juuri sellaista organisaatiossa, mikä tuottais lisäarvoa. Ja sitä kautta, jos sulla on governance ja ihmiset, verkostot kuvattuna ja foorumit on olemassa niin silloin sulla on tavallaan ne relevantit ihmiset määriteltynä datalle ja sitten tämmöisellä prototyypin validoinnillakin olemassa todennäköisemmin.”*

Datan ja siten tekoälyn hallinnointi on elintärkeää, osuu tekoälyhanke pienelle tai laajemmalle liiketoiminta-alueelle. Kokonaisvaltaisella datan hallinnoinnilla kuitenkin omalta osaltaan mahdollistetaan tekoälyn onnistunut skaalaus laajemmalle. *”Tässä ehkä sama, läpi ton koko kehityselinkaaren, niin mitä pienemmästä hankeesta on kysymys, niin jos tosi kapea ja siiloinen, niin se kakkonenkin saattais (datanhallinnan maturiteetissa datan hallinnoinnille) riittää, mut mitä kokonaisvaltaisempi, mitä enemmän erilaista dataa eri järjestelmistä se tekoälykehitysaihio koskee, sen tärkeemmäks nousee datan hallinnointikin, jotta pystyt muun muassa päätöksenteon ja sitten tietosuojan näkövinkkelistäkin varmistamaan, että oikeet tahot ovat tietoisia, että mitä tapahtuu. Eli kyl tässäkin mille tahansa tekoälykehitystä tekevälle se kolmostaso on sinne viimeistään tuotantoon siirtämiseen mennessä, mielellään prototyyppivaiheessa tarpeen.”*

Datan hallinnointiin kytkeytyy vahvasti myös tietosuojan ja muiden, esimerkiksi toimialakohtaisten säädösten mukaan toimimisen varmistaminen, erityisesti jos tekoälyhankkeessa aiotaan hyödyntää henkilötietoja tai muita liiketoimintakriittisiä tietoja. *”Sitten jos poimin seuraavaksi vaikka ton tietoturvan, jonka sisällä se tietosuoja on, niin vähän samanlaisen tietoturva-, tietosuojamaturiteetin tarttis tossa tekoälyprojektin aikana viimeistään asettua sinne nelostasolle ainakin. Optimoitu on semmosta toiveajattelua, mutta neloseen pitäis tähdätä, jotta ei tuu liian suuria riskejä tietoturva- ja tietosuojamielessä.”*

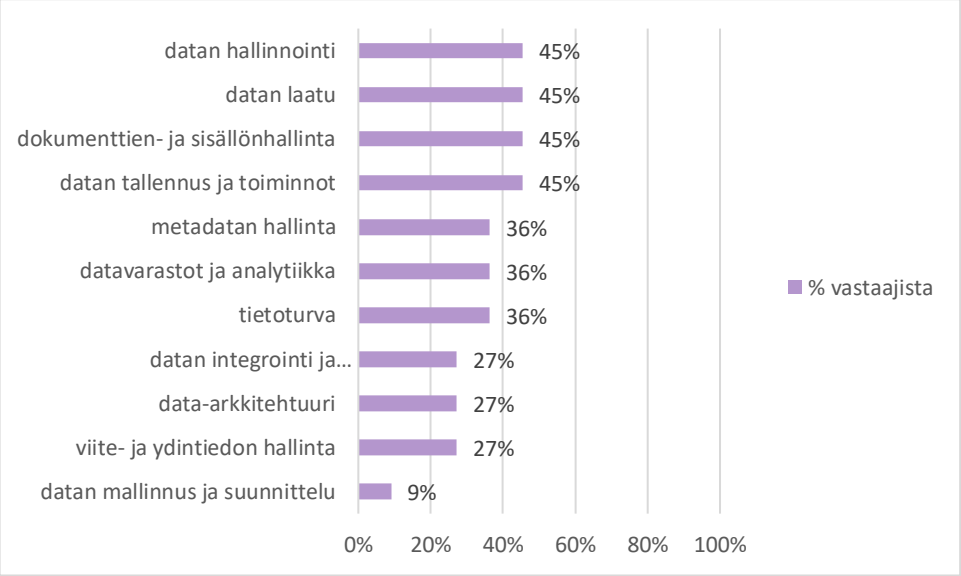
Datan laadun kehittämistä ja laadun monitorointikyvykkyyttä on kehitettävä läpi tekoälykehityksen ja tekoälyn elinkaaren. Liiketoimintahyötyä tuottavaan tuotantovaiheeseen päästään vasta, kun datan laadun hallintaan on kiinnitetty riittävällä vakavuusasteella huomiota. Haastatteluvastausten ja ideointityöpajan tulosten perusteella voidaan sanoa, että datan laatu on yksi niistä datanhallinnan osa-alueista, joita tulee kehittää ennakoivalta ja määritellyltä maturiteettitasolta korkeammalle kohti hallittua ja optimoitua maturiteettitasoa. *”Jossain prototyypin luonnin ja tuotantoon siirtämisen välissä olis pitäny tapahtua datan laatua, luotettavuutta parantava työ.”*

Kuten aiemmin ollaan todettu, hyvällä metadatan hallinnalla saadaan tuotettua kattava data katalogi organisaation datavarannoista, niiden ominaisuuksista ja hyödynnettävyydestä tekoälyn tarpeisiin. Kun datavarannot on määritelty, tekoälyn tuottama data on myös helposti upotettavissa osaksi standardidokumentaatiota. Näin tekoälyn tuotokset saadaan helpommin näkyville ja hyödynnettäväksi myös muualla organisaatiossa. *”Tekoäly tuottaa kanssa uutta dataa, niin sekin on standardisoitava ja tallennettava. Että sieltä tulee sitten metadatan hallinta. Prototyyppi, sillä pystytään validoimaan se, että pystyykö organisaatio esimerkiksi tuottaa heidän standardiinsa lisää kenttiä, vaikka sen mukaan, mitä se tekoälytuote tuottaa. Ja sitten kun pitää validoida jotenkin sen mallin toimintakin niin jos ei sinulla ole siihen määriteltyä dataa... Se data, jota sä käytät mallin validointiin, ei ole määritelty hyvin, niin miten sä ikinä pystyt määrittelemään, että onko se tekoälytuote hyvin toimiva vai huonosti toimiva.”*

Datanhallinnan osa-alueiden vaadittu maturiteetti tuotantovaiheessa

Vastaajat eivät painottaneet erityisesti yhden tietyn datanhallinnan osa-alueen kehittämistä, kun tekoälykehityksessä siirrytään viemään kehitetty tekoälyratkaisu tuotantoon (kuvio 7). Vastaajista 45 % piti sekä datan hallinnoinnin, datan laadun, dokumenttien- ja

sisällönhallinnan että datan tallennus toiminnot -osa-alueen riittävää maturiteettia yhtä tärkeänä tuotantoon siirto -vaiheessa.



Kuvio 7. Tärkeimmät datanhallinnan osa-alueet tekoälykehityksen tuotantoon siirtämisvaiheessa

Tekoäly ja sen hyödyntämä data tulee olla erittäin hallittua, jotta voidaan puhua tuotantokelpoisesta tekoälyratkaisusta. Mitä laajemmalle tekoälyn hyödyntämän datan lonkerot organisaatiossa ylettyvät eli mitä moninaisempaa ja erilaisista lähteistä tulevaa dataa hyödynnetään, sitä vaikeampaa on hallita tekoälyä ja sen käyttämää dataa. *”Mitä enemmän erilaista dataa eri järjestelmistä se tekoälykehitysaihio koskee, sen tärkeemmäks nousee datan hallinnointikin.”*

Datan hallinnointi -kyvykkyyttä on kehitettävä rinnakkain tekoälyn hyödyntämisen skaalaamisen kanssa, jotta mahdollisuus ohjata tekoälyn toimintaa säilyy liiketoiminnalla. *”(Tuotantoon siirtämisen vaiheessa) pitää olla selkeä roolitus. Täytyy olla hyvin hallussa semmoinen, että mitä tapahtuu, jos havaitaan esimerkiksi tekoälyn tuottamassa datassa laatuongelmia. Tällaiset roolitukset pitää olla selkeet siinä vaiheessa, kun viedään tuotantoon.”* Hallittu datan ja tekoälyn hallinnointi tarkoittaa sitä, että tekoälyratkaisun toimintaa ja sen taustalla olevassa datassa tapahtuvia muutoksia monitoroidaan ja auditoidaan säännöllisesti, jotta suunnittelemattomaan toimintaan ja muutoksiin päästään puuttumaan heti. *”Miten niitä pitäisi kehittää jatkossa eli siinä pitää olla omistajuus ja ymmärrys siitä, että mitä ollaan jo tehty. Ja jos ollaan tekoälyratkaisussa oikein modernilla puolella pitkässä kehityksen kaaressa, niin siellä varmaan tekoäly alkaa itsekin jo tavallaan kehittämään asioita ja tekee sitä kehitystä, niin kuinka sitten vastuulliset ihmiset pysyy sen kyydissä, jos mennään siihen, että kone ja algoritmit alkaa itsenäisesti kehittämään asioita*

(...). Varsinkin siinä vaiheessa, jos jotain menee pieleen. Että tavallaan sille on joku auditointi- ja monitorointikyvykkyys."

Jotta hallinnointia voidaan tehdä hallitusti, tarvitaan tekoälystä ja sen hyödyntämästä datasta dokumentoitua tietoa. *"Ehkä vielä siinä tuotantoon siirtämisessäkin sitä dokumentaatiota korostaisin. Että saadaan ymmärrystä, mitä on tehty, minkälainen putki siellä on, mistä me luetaan ne datat ja mitä me otetaan siihen malliin mukaan ja mihin me tallennetaan siihen malliin ja kun annetaan ennusteita, niin miten se toimii, mikä data meille tulee, mistä sisälle."*

Datan laatu on vastaajien mielestä yksi niistä datanhallinnan osa-alueista, joiden maturiteettivaatimus kasvaa, kun siirrytään tekoälykehityksessä tuotantovaiheeseen. *"Eli tuotantoon siirtämiseen mennessä datan laadun maturiteetin tarttis olla kuitenkin nelostasolla."* Datan hallinnointi on avain varmistamaan se tärkein eli tuotantokäytössä olevan tekoälyratkaisun käyttämän ja tuottaman datan laatu. *"Se on ihan älyttömän tärkeä koko ajan seurata sekä itse sitä datan laatua, että sitten sen AI:n kautta sitä dataa siellä lähdejärjestelmissä."* Tuotantoon siirretyn tekoälyratkaisun osalta on pitänyt varmistaa datan kattavuuden ja oikeamuotoisuuden lisäksi se, että data on myös sisällöltään oikeanlaista, jotta tekoäly pystyy tuottamaan valideja päätelmiä. *"Datan laatu on näistä kaikista a ja o. Miten itse näen, niin datan laatu sillai laajalla perspektiivillä, että datan laatuhan tarkoittaa, että se data on olemassa ja se on oikeamuotoista, mutta että se on myös oikeasisältöistä."*

Jotta datan laadun seuraaminen olisi hallittua, monitorointi on pystyttävä automatisoimaan, mitä laajemmalle skaalatusta tekoälystä on kyse. *"Riippuen use case:sta, että opetatko aina tietyin väliajoin mallia ja vertaat, että tuliko parempi jollakin setillä kuin edellisestä. (...) siinä jotakin automatiikkaa jo voisi ehkä olla, mikä tsekkaa datan laatua (...), että puuttuuko sinulla suuri osa datasta tai muuttuuko distribuutiot datassa kovasti (...). Niin ehkä toi datan laadun seuraaminen muodostuu minusta tärkeäksi siellä loppupäässä tai tuotannossa."*

Eri datalähteissä tapahtuvia datamuutoksia on myös kyettävä hallitusti ohjaamaan, monitoroimaan sekä linkittämään ne tekoälyn toimintaan mahdollisia ongelmatapauksia varten. *"Jos datan lähteet esimerkiksi vaihtuu tai jos jossain lähdejärjestelmässä tapahtuu muutoksia, niin silloin se metadatan hallinta on äärimmäisen keskeistä, koska muutenhan me ei tiedetä, että datan lähde muuttuu ja sitten voi käydä sillä tavalla, että meidän AI rupeaa sekoilemaan, koska se ei saakaan enää sitä dataa, mitä se luulee saavansa tai mitä sen pitäisi saada, vaan se saa jotain muuta dataa."* Tähän liittyy myös yksittäinen huomio

viite- ja ydintietojen laadusta, jos tekoäly hyödyntää tällaista dataa. *”Kun mennään tuotantoon, niin silloin myös viite- ja ydintiedot. Ne pitää olla tosi hyvässä kunnossa, että voidaan luottaa taustalla olevaan dataan.”*

Perusjärjestelmien vastuuhenkilöiden ja käyttäjien täytyy olla tietoisia siitä, miten järjestelmässä tehdyt muutokset vaikuttavat tekoälyn toimintaan. Jalkautetuilla data-standardeilla, dataprosesseilla ja toimintaohjeilla varmistetaan, että tekoälyn toimintaan haitallisesti vaikuttavia datamuutoksia ei tehdä perusjärjestelmiin alun perinkään. *”Jos lähdetään jotain isompia datan muutoksia tekemään, vaikka ihan tuolla perustiedoissa, niin onko sillä jotain mahdollisia vaikutuksia jo tehtyihin tekoälyratkaisuihin, joka voisi sitten aiheuttaa jotain muutoksia sinne. Tällaisia muutoksia varmaan täytyy pystyä manageeraamaan.”* Kun tekoäly hyödyntää useiden järjestelmien dataa, hallinta on oltava keskitettyä. Tässä siis korostuu jälleen kerran datan hallinnoinnin ja metadatan hallinnan tärkeys. Lisäksi data-arkkitehtuurin on kyettävä vastaamaan kasvaviin liiketoimintatarpeisiin. *”(Tuotannossa) ratkaisee se, että miten hyvin isot datamassat, kun siirrytään prototyypin kivan sample-otoksen ohi, niin se vaatii tosi paljon data-arkkitehtuuria.”* Data-arkkitehtuurin taso heijastuu datan laadun tasoon. *”Data-arkkitehtuuri täytyy olla tietyssä mielessä hallittuna ja mun mielestä taas toisaalta se hallittu data-arkkitehtuuri vaikuttaa datan laatuun.”*

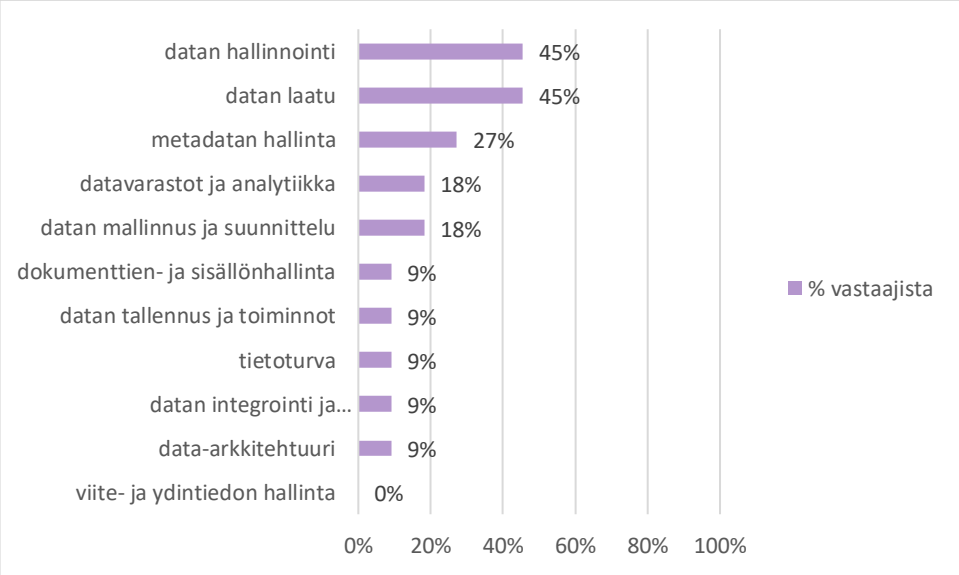
Tietoturvaan ja tietosuojaan liittyvien vaatimusten on toteuduttava tuotantokäyttöisissä tekoälyratkaisuissa. Siksi vaatimusten on oltava tiedossa jo AI-projektin alusta lähtien. *”Sitten tietoturva, että se on ihan hyvä olla hallussa jo alkupäässä, niin siihen ei käytännössä tarvitse enää tässä vaiheessa, mutta se on merkittävä osa myös tässä toki, että se data, esimerkiksi EU AI act:n mukaiset vaatimukset toteutuu viimeistään tuossa vaiheessa.”* Ennen tuotantoon siirtoa tekoälyn vaatimuksenmukainen toiminta on varmistettava erilaisten analyysien ja auditointien kautta. *”Tuotantoon siirtämiseen mennessä pitäis olla sitten kaikki riski-, uhka-analyysit tehty ja varmistettu, mitä tekoäly sille tiedonkäsittelylle tuo uutta tai muuttuvaa siihen nykytilaan, ennen tekoälyä, siihen tilanteeseen.”*

Tuotantoon siirtymisen vaiheessa tietysti myös tekoälyyn liittyvän datavarasto ja analytiikka -kyvykkyyden on oltava hallitulla maturiteettitasolla ja on tiedettävä, miten tekoäly hyödyntää dataa. *”Jos et ymmärrä nykyiselläänkään analytiikan keinoja, jos ne datavarastot ei oo kuosissa, niin en oikein tiedä, miten päästään mihinkään järkevään tekoälyn pyörittämiseen eli sen nostaisin ehkä sinne neloseen, että siinä ei oikein kolmonen taida piisata. Ja tuotantoon siirtämiseen mennessä.”* Datavarastot ja analytiikka ovat riippuvaisia datan integrointi ja yhteentoimivuus -kyvykkyydestä, jotta tuotannossa ei törmätä tiedonsiirtoon ja datan yhteentoimivuuteen liittyviin ongelmiin. *”Yhteentoimivuus ja integrointi*

kanssa. Se on iso asia. Se on jännä, miten tohon törmätään, että kun mennään tuotantoon isoissa asioissa ja nykyään tehdään asioita vaikka pilvessä. Ihan se, että kuinka paljon törmätään ulkoverkko-, sisäverkkoasioiden tiedonsiirto-ongelmiin ja muihin, mitkä kuulostaa ihan tyhmältä, mutta tällaisiakin on." Mitä laajempi tekoälyratkaisu on kyseessä, sitä vähemmän työn pitäisi sisältää manuaalisia vaiheita. *"Tääkin (datan integrointi ja yhteentoimivuus) on, että jonkun tarttee manuaalisesti syöttää tietoja tekoälyn suuntaan, jos nää ei oo kunnossa, niin nostetaan tääkin sinne nelostasolle ja tuotantoon siirtämiseen mennessä."*

Datanhallinnan osa-alueiden vaadittu maturiteetti AI:n elinkaaren hallinnassa

Tekoälykehityskaaren lopun eli AI:n elinkaaren hallinnan osalta vastaajat korostivat datan hallinnoinnin, datan laadun ja metadatan hallinnan tärkeyttä muihin datanhallinnan osa-alueisiin verrattuna (kuvio 8, 57). Vastaajista 45 % korosti datan hallinnoinnin ja datan laadun jatkuvaa kehittämistä sekä vastaajista 27 % korosti metadatanhallinnan maturiteetin kehittämisen tärkeyttä.



Kuvio 8. Tärkeimmät datanhallinnan osa-alueet tekoälykehityksen AI:n elinkaarivaiheessa

AI:n elinkaaren hallintaan kuuluu organisaation jatkuva strateginen kiinnostus dataa kohtaan sekä datan dokumentointi ja datan valjastaminen aiempaa laajemmin tekoälyn käyttöön. *"Liian pieni ymmärrys isoista massoista sitä perusdataa, mitä tullaan sitten loppujen lopuksi kuitenkin tuotantokäytössä hyödyntämään, niin se on yleensä liian pientä."* Tekoälyn skaalaamisen sekä metadatan hallinnan ja datan laadun kehittämisen tulee kulkea rinnakkain datan hallinnointi -rakenteiden ohjaamina, jotta organisaatio voi

ottaa oikean harppauksen tekoälykeskeiseen aikakauteen. Aluksi hyödyt näkyvät todennäköisemmin datan laadun paranemisena ja jatkossa monipuolisempina tekoälyn hyödyntämismahdollisuuksina. *”Mun kokemus on se, että AI on niin nuori, että käytännössä elinkaari on sitä, että koko aika datan laatu paranee ja paranee ja paranee.”*

Jotta datanhallinnan osa-alueet saadaan linkittymään toisiinsa ja niitä johdetaan keskitetysti ja hallitusti organisaation strategiset datalle asettamat tavoitteet ohjenuorinaan, datan hallinnoinnin on ohjattava kaikkea tekoälyyn liittyvää datanhallintaa. *”(Datan hallinnointi) on varmaan kaikkeen liittyvä. Tietysti voi ajatella, että se on tuolla elinkaaren hallintavaiheessa jotenkin erityisen tärkeätä ja keskiössä, mutta tavallaan vaikea ajatella, että noi muut asiat olisi kauhean hyvällä tolalla tai joku niistä voi olla, mutta että se kokonaisuus olisi kauhean hyvässä kunnossa, jollei governance olisi myös siellä paikallaan.”* Pysyvät datan hallinnoinnin rakenteet varmistavat tekoälyn ja siihen liittyvän datan jatkuvan monitoroinnin. Monitoroinnin ei pidä pysähtyä varsinaisen tekoälyhankeprojektin jalkautusvaiheen päätyttyä. *”Siinä (AI-elinkaaren hallinnassa) mun mielestä nousee tosi paljon governance sitten, että kun me huomataan asioita, mitkä toimii, mitkä ei, niin meillä pitää olla vaan mekanismit puuttua niihin. Että vaikka algoritmi ei tuota haluttua tulosta tai data ei ole laadultaan riittävää, niin meillä pitää olla mekanismit reagoida noihin ja laittaa ne kuntoon.”* Datan hallinnointirakenteen perustamista voidaan perustella myös lainsäädännöstä tulevien vaatimusten kautta. *”...pitää olla omistajuus kunnossa. Ja sitten tietysti kun ruvetaan puhumaan tietoturvasta, regulaatioista ja muusta tämmöisestä, niin sehän tietysti on aika oleellinen osa kanssa tekoälysovellutuksia, varsinkin jos EU:lta tulee vielä jotain pakottavaa lainsäädäntöä, jossa on aika merkittävät sanktiot, niin siinä alkaa olla hyvinkin suuri motivaatio saada datanhallinta kuntoon siellä taustalla. Puhumattakaan siitä sitten itse tekoälyn hallinnosta.”*

Tekoälyratkaisut hyötyvät parhaiten keskitetystä datavarastosta, jossa on monipuolista dataa. *”Se on haaste, että sitä dataa ei saada kerättyä yhteen paikkaan ja vielä ennen kaikkea jos se olisi sitten tämmöinen tuotantokäytössä oleva, jatkuvassa käytössä oleva tekoälyratkaisu niin silloin on tärkeätä, että se datavirta, että ne datat tulee sinne yhteen paikkaan ja se datavirta toimii, ei vaan että ne on kerättävissä, vaan se jatkuvasti pelittää niin että on vaikka tietovarasto, missä on mahdollisimman monipuolisesti sitä dataa käytettävissä.”*

Kun ensimmäinen tekoälysovellus on viety tuotantoon, palataan usein suunnittelupöydän äärelle ja ideoidaan sovellukselle jatkoa. Työ helpottuu, kun tekoälyn käyttämä data ja ylipäätään organisaatiolle tärkeä data kokonaisuudessaan on jo mallinnettu ja tietomalleja hallitaan ja päivitetään keskitetysti. Suunnittelussa ei siis tarvitse tällöin lähteä taas

tyhjältä pöydältä vaan ideoinnin tueksi löytyy ajantasaista tietoa data-arkkitehtuurin nykytilasta ja sitä kautta datan käyttömahdollisuuksista tekoälyn suuntaan. *”Totta kai elinkaaren hallinnan aikana korostuu myös se, että kun me kehitetään sovellusta eteenpäin, että elinkaaren hallinta ei ole pelkkää applications maintenance:a vaan se on mun mielestä myös jatkokehittämistä, että sitten kun mennään jatkokehittämiseen, niin totta kai se menee datan mallinnus ja suunnittelu -puolelle, että mitä uusia asioita sinne pitää ottaa, mitä tuoda mukaan.”*

Datanhallinnan maturiteetin analysointi tekoälykehityksen näkökulmasta

Datanhallinnan maturiteettianalyysia voidaan käyttää indikoimaan organisaation datanhallinnan kyvykkyyttä kehittää tuotantokelpoisia tekoälyratkaisuja. Analyysissa on huomioitava, miten eri datanhallinnan osa-alueiden aktiviteetit toteutuvat tosiasiallisesti käytännössä. Analyysia laatiessa on siis varmistettava, että vastaukset perustuvat todellisuuteen eikä siihen, miten datanhallinnan vain ajatellaan ja toivotaan toimivan. *”Kyllä se yleinen (datanhallinnan maturiteettiaste) on tärkeä, koska se indikoi niin voimakkaasti sitten jonkun yksittäisen datanhallinnan maturiteettia. (...) Se mitä siinä maturiteetissa pitää huomioida, ei se että onko sillä datalla omistaja, onko sillä vastuulliset, vaan kantaako ne sen vastuunsa, kantaako ne sen omistajuuden. (...) periaatteessa mä sanoisin että se pitää aina tehdä haastattelututkimuksena, haastatteluselvityksenä, jotta voidaan saada ihmisiltä ne kommentit siitä, että toimiiko se oikeasti vai eikö se toimi. Eli se pitää huomioida, että jotta meillä on sitä oikeata dataa sitä tekoälyä varten, me tiedetään, mitä se data oikeasti on, että se on oikeata dataa ja me tiedetään millä laatutasolla se on, niin me tarvitaan se, että meidän tiedonhallinnan kyvykkyys on käytössä eli meidän datan maturiteettiaste ei ole paperilla, meidän datan maturiteettiaste on todellinen, se on faktinen, se on toimiva se maturiteettitaso mikä me kuvitellaan, että meillä on. Ja se on mun mielestä hirveän vaikea asia, koska usein organisaation johdolla on aivan liian ruusuinen kuva siitä todellisudesta, että kuinka toimiva se datanhallinta oikeasti on. Kuvitellaan jotain. Meillä on nimetty nämä omistajat, mutta kun se nimeäminen ei tarkoita mitään, sen omistajan pitää oikeasti kantaa se vastuu.”*

Ennen datanhallinnan maturiteettianalyysia on kuitenkin syytä selvittää organisaation tavoitteet, jotka se haluaa saavuttaa tekoälykehityksen kautta, ja skaalata maturiteettianalyysi sen mukaan. Datanhallinnan maturiteettianalyysi voidaan siis kohdistaa koko organisaation sijaan myös esimerkiksi yhdelle liiketoiminta-alueelle tai yhteen liiketoimintaprosessiin. *”Tekoälyäkin on niin hirveän monenlaista. Sitten taas just se, että halutaanko sitä käyttää johonkin, miten nyt vaikka KELA tai jotku aikuisopintotuet katsotaan, et onko ne kyllä/ei -juttuja, niin eihän siinä tarvitse olla maturiteetin hirveän korkealla, että jotenkin*

se pitäisi määritellä, että mihin se firma, mitä se tavoittelee ja mitä se haluaa ja sen kautta katsoa, että mikä on se tavoiteltava maturiteettitaso, koska eihän kaikkien firmojen tarvitse päästä mihinkään maailman kypsimpään vaiheeseen. Joskus semmoinen keskivaihekin riittää, mut sit ei voida taas tavoitella niitä ihan vimosen päälle olevia tekoälyratkaisuja. (...) monet aina puhuvat siitä, että pitää mennä maturiteeteissa ihan loppuun ja päätyyn asti ja sitten kun ollaan siellä, niin taivas repeää ja kaikki onnistuu. Mutta että kyllä se vähempikin maturiteetti riittää, mutta silloin ei vaan voi tavoitella taivaita. Eli sen pitäisi olla aina suhteessa siihen mitä halutaan, niin sen tavoiteltavan maturiteettiasteenkin."

Datanhallinnan maturiteetin parantamiseen liittyy varsinaisten työkalujen ja säännöstöjen kehittämisen lisäksi tietysti myös ihmisten toiminnan ja prosessien kehittäminen. *"Sinällään ensin kannattaakin tehdä semmoisia jotain hyvin alkeellisia tekoälyratkaisuja ja katsoa miten se toimii, että samalla kun opettaa niitä koneita, niin opettaa myös ihmisiä toimimaan. Että onhan tämä myös hirveän iso asia, joka muuttaa työtä...(...) pitää myös opettaa se ihmisen ja koneen välinen yhteistyö, niin sitä kautta ku tehdään pala palalta, kasvatetaan maturiteettia ja samalla vaikeutetaan sitä, niin se on ehkä se toimiva, ainoa toimiva tapa."*

Tekoälykehityksessä on huomioitava myös auditointikyvykkyyden mahdollistaminen, mikä tulee vastaan viimeistään tuotannossa olevan tekoälyratkaisun kohdalla. *"Siis maturiteettia voidaan mitata hyvinkin yksityiskohtaisesti, että mä puhuisin tästä tekoälyyn liittyvästä maturiteetista, mä menisin pikkuisen lähemmäksi auditoinnin suuntaan jo sen takia, että siellä on asetuksia ja muita ja kun mennään auditoinnin suuntaan, niin meillä pitäisi olla lista asioista, joita me myös auditoidaan, et ehkä meille tulee myös minimivaatimuksia mitä pitää tsekata. Governanceahan voitaisiin miettiä ylipäätänsä tavallaan helposti yksinkertaistetulla tavalla, että onko governancea olemassakaan ja onko se governance semmoinen, että se on tehty kerran ja kehittyykö se. Onko se semmoinen, että ihmiset jollain lailla toimii sen mukaan ja miten se on. Mutta sitten voitaisiin ajatella, että kyllä mun mielestä data governancea vaikka DAMA-pyörän kannalta, niin pitäisi katsoa jo niitä osa-alueita ja mun mielestä tekoäly tuo tietysti vielä sitä omaa tarkkuustasoa. Ehkä tekoäly tuo sen, että meille tulee varmaan niin kun jos ajatellaan, että ykköstaso on governance ja kakkostaso on datan laatu ja kolmostaso on laadun juttuja, niin varmaan laatuun vaikuttavia attribuutteja, niin varmaan vielä nelostaso löytyy tekoälyssä, joka on ehkä sitten semmoinen tosiaan, vaikka vähimmäisvaatimukset on kunnossa ja on katsottu."* Auditointikyvykkyyden rakentamisessa tarvitaan muun muassa kommunikointia tietoturva- ja tietosuoja-asiantuntijoiden kanssa. *"Holistisuus on niin tärkeetä. (...) ...mitä se tekoäly tekee sille datalle, niin tavallaan se näkyvyys ja ymmärrys siitä, että mitä tässä koko prosessissa tapahtuu, niin ne jotka on esimerkiksi tekoälyn ammattilaisia, niin ne*

varmasti tarvitsee tietosuojamielessä sparrausapua tietosuojaosaajilta ja tietoturvaosaajilta ja sitä kommunikaatiota auttaa tosi paljon tai se on ihan elimellisen tärkeää, että sitten henkilöt, jotka ovat tekoälyhankkeessa mukana, ne oikeesti pystyy edustamaan sitä näille vaikka tietoturvan ja tietosuojan sparraajille ja arvioijille, että mitä tässä oikein tapahtuu että mikä on se tietosisältö. (...) ...että ymmärrys ja dokumentaatio pysyy ajantasalla, jotta mahdollisimman ketterästi pystytään viemään sitä tekoälykyvykkyyttä eteenpäin, niin se jotenkin pitäis saada haltuun."

Jos organisaatiossa on jo tehty tekoälyhankkeita, datanhallinnan maturiteettia arvioitaessa tulee kiinnittää huomiota myös edellisten hankkeiden sujuvuuteen*. "Ehkä sen avainkysymykset on just samat, mitkä ylipäätään datanhallinnan, että onko olemassa jotain visiota, onko olemassa omistajia, onko olemassa säännöstöjä ja kuin proaktiivista versus reaktiivista se on ja sitten siinä alkuvaiheen kysymysten jälkeen varmaan päästään siihen, että kuinka helppoa tai vaikeata esimerkiksi tekoälyprojektit tai -ideoinnit on olleet että tukeeko se perusdata. Kuinka helppoa, vaikeaa on ollut luoda tekoälyratkaisuja. Ja niin edespäin, että kyl se tekoäly varmasti saataisiin siihen maturiteettimalliin jotenkin mukaan tämmöiseen yleiseen. Ja sitä kautta sitten varmaan tekoälyssä on varmasti paljon eri vaiheita. On ihan semmosia perus simppelijuttuja, mitä varmaan lähdetään pilotoimaan, mutta sitten mitä enemmän ja syvällisemmin ja laajemmin sitä lähdetään hyödyntämään, niin ehkä sieltä tulee sitten niitä erilaisia maturiteettipisteitä. Että kuinka automaattisesti tekoäly pystyy myös ideoimaan asioita versus että pitääkö siinä olla paljon ihmistyövoimaa ja niin edespäin."* Datanhallinnan maturiteetti näkyy esimerkiksi siinä, miten datan laatua jo seurataan ja miten dataa arvostetaan. *"Mä sanoisin, että ihan johdannaisena mun aikaisemmista vastauksista niin kiinnitetäänkö datan laatuun huomiota, seurataanko sitä datan laatua, niin se olisi yksi semmoinen, mikä indikoisi, että arvostetaan sitä dataa, nähdään, että se on tärkeätä, että se on kunnossa. Maturiteetti on korkeampi, jos tosiaan näin tehdään ja mahdollisimman pitkäjänteisesti tuota ja ennaltaehkäisevästi jo pyritään ehkäisemään dataan liittyviä virheitä. Esimerkiksi virhesyöttöjen mahdollisuuksia erilaisissa järjestelmissä, niin tämmöinen voisi indikoida sitä, että jo ennaltaehkäisevästi pyritään vaikuttamaan siihen, että data pysyy hyvälaatuisena."*

Datanhallinnan korkeampi maturiteettitaso riippuu voimakkaasti organisaation tärkeimpien osasten eli ihmisten suhtautumisesta dataan ja datanhallintaan. *"...sitten mä miettisin sitä, että voisiko siinä olla myös jotain pehmeitä kriteereitä, millä sitä arvioidaan, mitkä liittyy siihen kulttuuriin. (...) mutta ennen kaikkea pitää olla ihmisillä semmoinen käsitys, että datanhallinta on tärkeätä ja osataan sanoa, että miksi se on tärkeätä ja oikeasti nähdään datanhallinta osana sitä työtä. Vaikkakin data palvelee varsinaista tekemistä, niin siltikin se data on niin keskeisessä roolissa, että se pitää nähdä osana sitä työtä ja toimenkuvaa,*

että siitä pidetään huolta ja mietitään datan laatuun liittyviä kysymyksiä. Jos jotenkin pääsisi tässä maturiteetin arvioinnissa käsiksi siihen, että mitä ihmiset ajattelee siitä datasta ja sen hallinnan tärkeydestä ihan oikeasti ja miten ne sitte arjessaan sitä huomioi. Ne on varmaan yllättävän pieniä juttuja, arkisia juttuja loppujen lopuksi. Ei se ole mitään niin ihmeellistä, mutta sitä jotenkin kyllä varmasti haluaisin myös lähestyä ideaalisesti tämmöisessä datamaturiteetin arvioinnissa."

5 Tulokset

Tässä opinnäytetyössä asetettiin yhteensä kolme tutkimuskysymystä datanhallinnasta ja sen suhteesta tekoälykehitykseen. Yhtenä tutkimuskysymyksenä opinnäytetyössä pyrittiin vastaamaan siihen, mitä hyvä datanhallinta tarkoittaa käytännön tasolla ja miten se heijastuu organisaatioiden toimintaan. Haastateltavien mukaan hyvä datanhallinta ilmenee suoraviivaisena työnä ja mahdollisuutena hyödyntää dataa mitä moninaisemmin tavoin liiketoiminnan tarkoituksiin. On kuitenkin ymmärrettävä, että tähän pisteeseen ei päästä, ellei organisaatiossa oikeasti ja laajasti anneta datalle ja datanhallinnalle sen tarvitsemaa arvostusta. Hyvä datanhallinta on sidoksissa organisaation toimintakulttuuriin, koska datanhallinta lähtee ihmisten datalähtöisestä ymmärryksestä, arvostuksesta ja toiminnasta, mikä taas näkyy tehokkaina prosesseina ja oikein valjastetusta teknologiasta lähtöisin olevina liiketoimintahyötyinä. Haastateltavat kuvasivat yhteneväisesti myös sitä, miten vastaavasti huono datanhallinta ilmenee organisaatioissa. Tällöin datapääoman potentiaali jätetään kartoittamatta ja hyödyntämättä, mutta lisäksi riskeerataan kaikkien datapohjaisten hankkeiden eteneminen, kun data ei ole luotettavaa, suojattua eikä helposti hyödynnettävissä, jos datasta on ylipäätään tietoa saatavillakaan.

Toisen tutkimuskysymyksen osalta opinnäytetyön tavoitteena oli vastata tarkemmin siihen, millainen rooli datahallinnalla on tekoälykehityksessä. Kaikki haastateltavat näkivät datanhallinnan tason vaikuttavan tekoälykehitykseen. Hyvä datanhallinta nähdään kilpailuvalttina silloin, kun tekoälyratkaisu siirretään tuotantoon. Osa haastateltavista huomautti, että monet organisaatiot eivät kuitenkaan anna datanhallinnalle sen ansaitsemaa huomiota eivätkä ymmärrä dataansa, vaikka he samalla tiedostavat tekoälyn kasvavan merkityksen. Hyvä datanhallinta kuitenkin tukee vahvasti tekoälykehityksen onnistumista, koska tekoäly on riippuvainen sille syötetystä datasta. Jos datan laatu on huonoa ja datanhallinta datan laadun ennaltaehkäiseväksi parantamiseksi olematonta, tekoälyasiantuntijoiden arvokas osaaminen valuu datan laadun kehittämiseen liiketoimintahyötyä tuovan tekoälykehityksen sijaan ja hankkeet venyvät. Huono datan laatu voi johtaa myös hankkeen kaatumiseen tai tuotantoon lisätyn ratkaisun lainsäädännöllisestäkin näkökulmasta kyseenalaiseen toimintaan, johon on mahdoton puuttua ilman ihmislähtöistä kontrollia ja datanhallinnan hyviä käytäntöjä. Haasteltavien mukaan ihmisillä on aina säilyttävä lopullinen vastuu sekä kyky tehdä päätöksiä ja puuttua tekoälyn toimintaan, kun se on tarpeellista. Kyky hallinnoida tekoälyä säilyy, kun datanhallinnan prosessit ja dokumentaatio tukevat sitä. Hyvät datanhallinnan käytännöt tukevat myös tekoälyn tuottaman datan vaatimuksenmukaisuutta ja hyödynnettävyyttä myös muualla organisaatiossa.

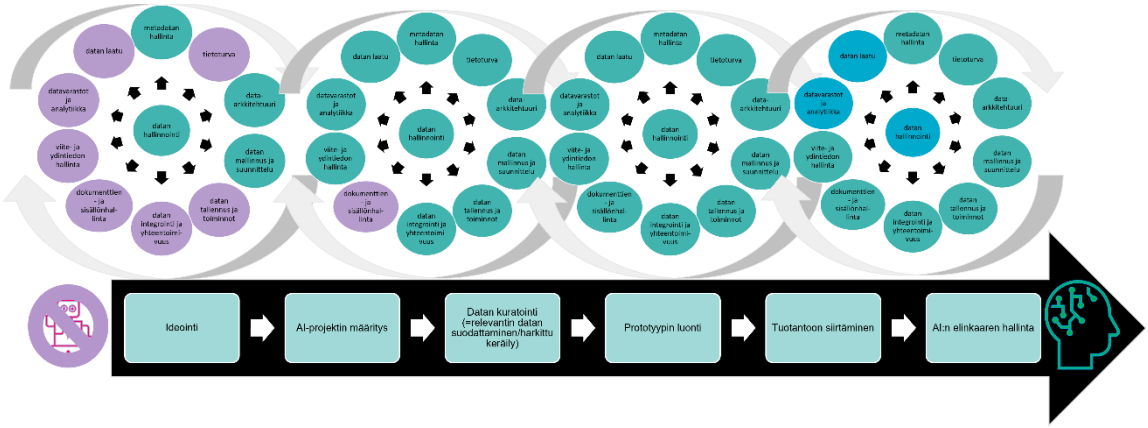
Tarkentavana kysymyksenä haastateltavilta kysyttiin, mitä lyhyen ja pitkän aikavälin hyötyjä he näkevät, mitä hyvästä datanhallinnasta voi koitua organisaatioiden tekoälykehitykselle. Lyhyellä aikavälillä datanhallinnan kehitys näkyy parantuneena datan saatavuutena ja laatuna, mikä taas näkyy siinä, että tekoälykehitys ei pysähdy dataan liittyviin ongelmiin ja näin saadaan nopeampaa arvontuotantoa. Lisäksi esimerkiksi datan hallinnoinnilla voidaan karsia päällekkäistä kehitystyötä ja identifioida paremmin liiketoimintaa parhaiten hyödyntäviä hankkeita. Pitkällä aikavälillä organisaation datavarantojen koko potentiaali voidaan hyödyntää hyvän datanhallinnan ansiosta, jolloin liiketoiminta saa arvoa ja arvonnousua. Paremman datan laadun kautta myös tekoälyn hyödyntämismahdollisuudet laajentuvat sekä organisaation sisällä että mahdollisesti myös organisaation ulkopuolelle eri tahojen kanssa käytävän yhteistyön kautta.

Opinnäytetyön kolmantena tutkimuskysymyksenä pyrittiin vastaamaan siihen, miten tekoälykehitykseen liittyvän datanhallinnan maturiteettia voidaan arvioida. Tutkimuksen tulosten perusteella voidaan sanoa, että tekoälykehitykseen lähtevän organisaation datanhallinnan maturiteettia voidaan arvioida hyödyntäen jo olemassa olevia datanhallinnan maturiteettimalleja, mutta mallissa tulee lisäksi korostaa tekoälykehityksessä merkittävimpiä datanhallinnan asioita ja huomioida, että datanhallinnan maturiteetilta vaaditaan lähtökohtaisesti jo enemmän verrattuna perinteisiin organisaatioihin. Tekoälykehitykseen ei siis voida järkevästi ja turvallisesti lähteä olemattomalla eli nollatason datanhallinnan maturiteettitasolla. Siksi lopulliseen tuotokseen eli painotettuun datahallinnan maturiteettimalliin ei otettu mukaan kyseistä nollatasoa vaan malli koostuu asteikosta yhdestä viiteen, jossa taso yksi on alustava tai tapauskohtainen, taso kaksi on toistettavissa ja reaktiivinen, taso kolme on ennakoiva ja määritelty, taso neljä on hallittu ja taso viisi on optimoitu ja tehokas datanhallinnan maturiteettitaso.

AI-valmiuteen eli tuotantokelpoiseen tekoälyratkaisuun tähtäävän organisaation datanhallinnan tavoitetilamaturiteetti määräytyy suoraan tekoälykehityksen datanhallinnalle asettamien vaatimusten kautta. Jotta pystyttiin vastaamaan tarkemmin siihen, millä kehitysprioriteetilla kutakin datanhallinnan osa-aluetta tulisi kehittää ja mikä datanhallinnan maturiteettitaso vaaditaan kultakin osa-alueelta AI-valmiuteen, haastateltavilta kysyttiin, mitkä heidän mielestään ovat kriittisimmät datanhallinnan osa-alueet kutakin tekoälykehityksen vaihetta kohden ja millainen maturiteettitaso näille vaaditaan missäkin vaiheessa, jotta voidaan sujuvasti kulkea kohti AI-valmiutta.

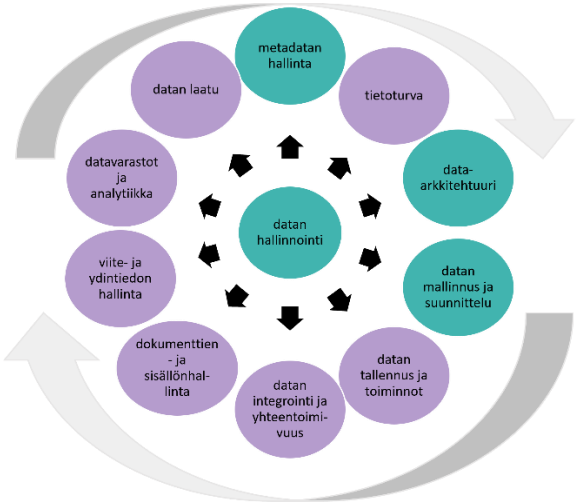
Havaintojen perusteella voidaan sanoa, että datanhallinnan eri osa-alueiden kehittämistarpeet painottuvat eri kohdissa tekoälykehitystä, kun organisaatiossa tavoitellaan AI-valmiutta eli tuontantokelpoista tekoälyratkaisua (kuva 6, 64). Riippuen

siitä, millä laajuudella organisaatio lähtee hyödyntämään tekoälyä, datanhallinnan maturiteettivaatimus kohdistuu valitun tekoälyratkaisun hyödyntämis- ja vaikutusalueen datanhallinnan tasoon.



Kuva 6. AI-valmiin organisaation datanhallinnan osa-alueiden painopisteet tekoälykehityksessä (mukaillen Sebastian-Coleman 2018, Coveyduc & Anderson 2020. Etelälahti 2021)

Haastattelujen ja ideointityöpajan perusteella tekoälykehityksen ideointivaiheeseen lähdettäessä on tärkeintä, että seuraavat datanhallinnan osa-alueet on jo kehitetty maturiteettitasolle 2 tai 3: metadatan hallinta, data-arkkitehtuuri, datan mallinnus ja suunnittelu sekä datan hallinnointi (kuva 7). Kuten aiemmin on todettu, olematon maturiteettitaso ei ole kuitenkaan riittävä muidenkaan datanhallinnan osa-alueiden osalta vaan kaikkien osalta tarvitaan vähintään ymmärrystä niiden tilasta, jolloin valitun kehittämisalueen jälkeen nähdään, mitkä muut alueet vaativat datanhallinnan osalta kehitystyötä.



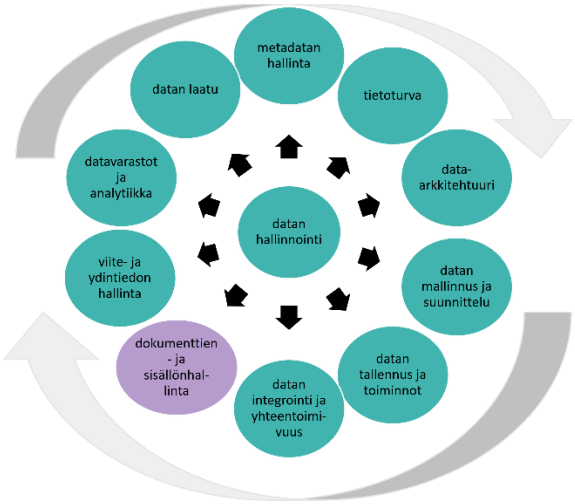
Kuva 7. Datanhallinnan osa-alueiden painopisteet tekoälykehityksen ideointivaiheessa (mukaillen Sebastian-Coleman 2018. Etelälahti 2021)

Tekoälykehityksessä data on yksi huomioitavista resursseista jo ideointivaiheesta lähtien. Haastateltavat korostivat erityisesti metadanhallinnan kriittisyyttä tekoälykehityksen alusta lähtien. Jotta ideointi onnistuu, tarvitaan tietoa datasta eli metadatanhallinta täytyy olla tältä osin tarpeeksi kattavalla tasolla. Dataa on muutoin mahdoton hallita saati ideoida sen pohjalta, jos ideointia joudutaan tekemään IT-vetoisesti tietokannoista käsin. Jos liiketoiminnalle ei ole saatavilla tietoa datasta sille ymmärrettävässä muodossa, niin riski kasvaa sille, että tekoälyratkaisu ei tule palvelemaan liiketoimintaa parhaalla mahdollisella tavalla.

Suunnittelun tukena hyödynnettävät ja luotavat tietomallit ovat myös datasta kertovaa metatietoa. Data-arkkitehtuurin sekä datan mallinnuksen ja suunnittelun kyvykkyydet luovat uuden suunnittelulle nykytilaan perustuvan pohjan datarakenteista, joita voidaan lähteä kehittämään kohti tavoitetilaa. Ilman olemassa olevia kuvauksia hankkeet venyvät ja kiireen pakottamana kuvaukset tehdään virhealttiisti. Lisäksi ideointivaiheeseen lähdettäessä organisaatiossa tulisi olla rakenneaihio datan hallinnoinnille, jotta varmistutaan siitä, että kaikki datanhallinnan osa-alueet palvelevat strategisia päämääriä tekoälyn suhteen ja että dataa hyödyntävä kehitystoiminta on perusteltu myös oikeutuksen ja luvallisuuden osalta. Datan hallinnoinnin kautta tehdään datan hallintatyön priorisointia ja valvotaan kehitystyön edistymistä. Datan hallinnoinnilla varmistetaan, että tekoälyn toiminnasta vastuussa olevat henkilöt saavat kaiken tarvitsemansa informaation tekoälyn toiminnan oikeanmukaista hallinnointia varten ja että kehityksen suunta on organisaation strategian mukainen ilman tarpeettomia riskejä. Pitkällä tähtäimellä jalkautetun datan hallinnoinnin prosessit, käytännöt ja säännöt ovat hyödynnettävissä laajemmin, kun tekoälyä skaalataan.

AI-projektin määritysvaiheeseen mennessä kaikki datanhallinnan osa-alueet tulisi olla kehitetty tasolle 3 eli ennakoivalle ja määritellylle datanhallinnan maturiteettitasolle lukuun ottamatta dokumenttien- ja sisällönhallinta -osa-aluetta (kuva 8, 66). Tämänkin osalta vaatimustaso on toki isompi, jos tekoälyratkaisuun on tarkoitus syöttää strukturoimatonta dataa. Vastauksissa nostettiin esille se, että datan erottelu eri osa-alueiksi eli sekä

viite- ja ydintiedoksi että strukturoimattomaksi tiedoksi ei ole välttämättä järkevää, koska kaiken tekoälyn hyödyntämän datan hallinta ja elinkaaren tuntemus on oltava riittävällä maturiteettitasolla.



Kuva 8. Datanhallinnan osa-alueiden painopisteet AI-projektin määritys- ja kuratointivaiheessa (mukaillen Sebastian-Coleman 2018. Etelälahti 2021)

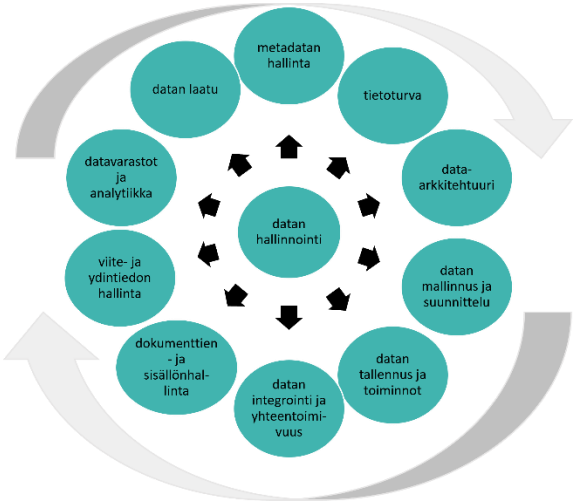
AI-projektin määritysvaiheessa pitää selvittää, sisältääkö datan hyödyntäminen arkaluonteisen datan hyödyntämistä. Selvityksessä hyödynnetään jo aiemmin kehitetyn metadatanhallinnan dokumentaatiota. Lisäksi arkaluonteisen datan hyödyntämistä varten muun muassa pääsyn- ja oikeuksienhallinnan tulee olla kunnossa. Hyvällä tietoturvalla ja tietosuojalla varmistetaan, että kehitettävän tekoälyratkaisun toiminta on luottamuksenarvoista. Tekoälykehityksessä hyödynnettävän datan laatu on kyettävä selvittämään sujuvasti ja kattavasti. Datan elinkaari tallennuspaikkoineen on siis oltava tiedossa. Näin myös tekoälyratkaisun kehityksen jälkeen osataan huomioida, mitkä datamuutokset vaikuttavat tekoälyyn millä eri tavoin ja millaisia muutoksia ei ole hyväksyttyä tehdä. Näkyvyys tekoälykehityksessä hyödynnettävän datan elinkaareen varmistaa myös tekoälyn läpinäkyvän toiminnan.

Vaatimukset datan laadulle tulevat tekoälykehityksestä ja sisältävät yleensä perinteisten datan laadun ulottuvuuksien lisäksi vaatimuksia datan saatavuudelle, kattavuudelle ja aikajanan pituudelle, jolla dataa on kerätty. Ei myöskään riitä, että tekoälykehitykseen tarvittavan datan laatua kehitetään pelkästään tekoälyhankkeen yhteydessä vaan datan laatua tulee pyrkiä parantamaan lähdejärjestelmistä lähtien, jotta data palvelee tekoälyratkaisua kestävästi pitkällä tähtäimellä. Jos nykyiset datan tallennusratkaisut eivät palvele tekoälykehitystä, dataa kannattaa keskittää, jotta se on parhaalla mahdollisella tavalla hyödynnettävissä. Valitulta ratkaisulta tarvitaan riittäviä analyyttisiä kyvykkyyksiä. Lisäksi datan integraatiokyvykkyyden ja yhteentoimivuuden on oltava saumatonta, jotta tekoäly

saa oikeamuotoista dataa oikea-aikaisesti. AI-projektin määritysvaiheessa tarvitaan päätöksiä ja toimenpiteitä näihin kaikkiin liittyen.

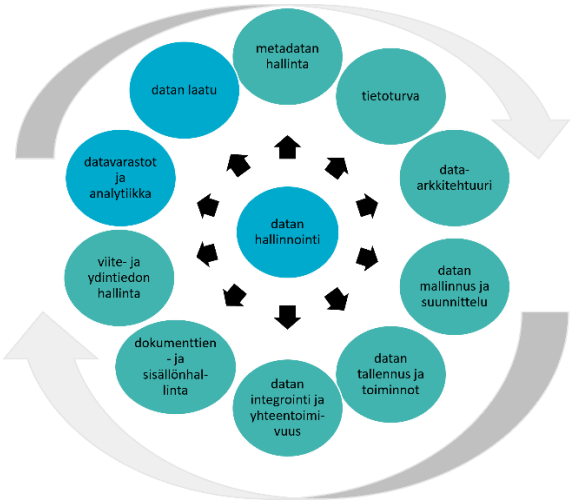
Datan kuratointi -vaiheessa datanhallinnan osa-alueiden painopisteissä ei tapahdu muutoksia AI-projektin määritysvaiheeseen verrattuna vaan edelleen kaikilta dokumenttien ja sisällönhallintaa lukuun ottamatta vaaditaan tason 3 maturiteettia (kuva 8). Riittävään dokumentointiin on tässä vaiheessa kuitenkin alettava kiinnittää enemmän huomiota, kun dataa kerätään tekoälykehitystä varten, jotta tekoälyn hallittavuus ja skaalattavuus säilyvät. AI-projektin määrityksessä määritellyt datanhallinnan toimenpiteet näkyvät muun muassa kehittyneenä integraationkyvykkyyden ja datan yhteentoimivuuden paranemisena sekä selkeinä prosesseina siitä, kuka, miksi ja milloin dataa kerätään tekoälyä varten. Datan kuratointi -vaiheessa tiettyjä osa-alueita aletaan kuitenkin kehittää jo kohti seuraavaa maturiteettitasoa. Kun datan laatuun ja laadun ymmärrykseen on kiinnitetty huomiota jo aiemmin, tekoälyhanke ei kaadu datan laadusta johtuviin ongelmiin. Datan kuratoinnin kautta datan laatu kehittyy edelleen.

Prototyypin luontivaiheessa tarvitaan lopulta myös dokumenttien ja sisällönhallinnan osalta johdonmukaista toimintaa eli tason 3 maturiteettia, jotta tekoälymalleista sekä niiden toiminnasta ja niiden hyödyntämästä datasta on saatavilla kattavat dokumentaatiot tekoälyn hallittavuutta ja skaalattavuutta ajatellen (kuva 9). Tiedot tekoälyn tuottamasta datasta upotetaan osaksi metadatanhallintaa eli tyypillisemmin osaksi data katalogia. Jo olemassa oleva datan hallinnointirakenne ulotetaan tekoälyn tuottamaan uuteen dataan ja varmistetaan, että tekoälyn toiminta täyttää tietoturvan ja tietosuojan antamat reunaehdot. Datan laadun osalta kehitetään monitorointikyvykkyyttä.



Kuva 9. Datanhallinnan osa-alueiden painopisteet prototyypin luonti -vaiheessa (mukaillen Sebastian-Coleman 2018. Etelälahti 2021)

Kun tekoälyratkaisu siirretään tuotantoon, tekoälyn ja sen hyödyntämän datan täytyy olla erittäin hallittuja. Haastattelujen ja ideointityöpajan perusteella AI-valmius tarkoittaa sitä, että datan hallinnoinnin, datan laadun sekä datavarastojen ja analytiikan maturiteetin täytyy olla tasolla 4 eli toiminnan pitää olla hallittua (kuva 10). Näiden osa-alueiden osalta nähtiin olevan realistista, relevanttia ja aidosti hyödyllistä vaatia maturiteetin jatkokehittämistä kolmatta maturiteettitasoa pidemmälle tekoälykehityksen puitteissa. Muiden datanhallinnan osa-alueiden osalta datanhallinnan maturiteetti voi jäädä tasolle 3, jolloin toiminta on kuitenkin jo ennakoivaa ja määriteltyä. Näiden osa-alueiden osalta maturiteetin jatkokehittämisessä nähtiin myös hyötyjä, mutta ne eivät ole edellytys AI-valmiudelle.



Kuva 10. Datanhallinnan osa-alueiden painopisteet tuotantoon siirtovaiheessa ja AI:n elinkaaren hallinnassa (mukaillen Sebastian-Coleman 2018. Etelälahti 2021)

Hallinnoinnin tärkeys korostuu, mitä laajempi tekoälyn toiminta- ja vaikutusalue on. Datan hallinnointia onkin kehitettävä kohti keskitettyä mallia, joka tukee skaalautuvaa tekoälyn hyödyntämistä parhaalla mahdollisella tavalla. Johdonmukaisuutta tarvitaan myös datan hallinnoinnin alle kuuluvan datan laadun ja siihen liittyvien riskien hallinnalta. Datan laadussa, sisältäen esimerkiksi datan oikeellisuuden, kattavuuden ja saatavuuden, on täytynyt tapahtua mitattavaa kehitystä. Jotta myös datan laadun monitorointi olisi skaalattavissa, monitorointia on automatisoitava. Datavarastoilta ja analytiikalta vaaditaan siis korkeampaa kyvykkyyttä vastata tekoälyn tarpeisiin.

Kun ensimmäinen tekoälyratkaisu siirretään tuotantoon, organisaatio siirtyy AI-matkallaan AI:n elinkaaren hallintaan. Koko organisaation on ymmärrettävä datan merkitys tekoälykehityksessä ja toimittava sen mukaisesti. Datan laatu paranee ja tekoälyn hyödyntämismadollisuudet laajenevat, kun tekoälyä kehitetään datan hallinnointi -rakenteiden ja hyvien datanhallinnan käytäntöjen ohjaamina, unohtamatta tekoälyratkaisujen auditointikyvykkyyden mahdollistamista. Tekoäly ja sen hyödyntämä data vaatii jatkuvaa monitorointia,

mutta ihmislähtöistä hallinnointia, jotta tekoälyn toiminta vastaa ihmisen tekoälylle asettamia tarpeita nyt ja tulevaisuudessa.

5.1 Tuotos: painotettu datanhallinnan maturiteettimalli

Opinnäytetyön aineiston perusteella ei voitu määrittää AI-matkalle valmiin organisaation datanhallinnan osa-alueiden vähimmäismaturiteettitasoa vastausten suuren hajonnan vuoksi. Haastatteluissa ja ideointityöpajassa oltiin kuitenkin yhtä mieltä siitä, että tekoälykehitykseen ei voida lähteä, jos datanhallinnan maturiteetti on olematonta. Aineiston perusteella voitiin kuitenkin määrittää AI-valmiin organisaation eri datanhallinnan osa-alueiden tavoitematuriteettitaso, joka on esitetty tutkakaaviossa (kuvio 9). Tässä opinnäytetyössä AI-valmiilla organisaatiolla tarkoitetaan datanhallinnan näkökulmasta tuotantokelpoisen tekoälyratkaisun kehittänyttä organisaatiota. Tuotantokelpoisuuden varmistamiseksi tarvitaan ennakoivaa ja määriteltyä toimintaa metadatan hallinnan, tietoturvan, data-arkkitehtuurin, datan mallinnuksen ja suunnittelun, datan tallennus ja toimintojen, datan integrointi ja yhteensopivuuden, dokumenttien ja sisällönhallinnan sekä viite- ja ydintiedon hallinnan osalta. Lisäksi tarvitaan hallittua toimintaa datan hallinnoinnin, datan laadun sekä datavarastot ja analytiikan osalta.

Kuvio 9. AI-valmiin organisaation datanhallinnan maturiteetti.

AI-valmiin organisaation datanhallinnan osa-alueiden tavoitetilan perusteella luotiin tekoälykehityksen näkökulmasta painotettu datanhallinnan maturiteettimalli (liite 3). Taulukon koon vuoksi maturiteettimalli on kokonaisuudessaan luettavissa liitteestä kolme, mutta kuvassa 11 on nähtävissä osa mallista esitettynä esimerkinomaisesti. Taulukko-muotoisen maturiteettimallin riveillä on kaikki datanhallinnan osa-alueet ja sarakkeissa datanhallinnan maturiteettiasteikko yhdestä viiteen. Mallista on luettavissa jokaisen maturiteettitason määritelmä toteamuksina kunkin datanhallinnan osa-alueen osalta. Toteamuksien perusteella organisaatiot voivat hahmottaa datanhallintansa nykyistä maturiteettitasoa. Maturiteettimalliin on liitetty tutkakaaviossa esitetty tavoitetila AI-valmiin organisaation maturiteettitasolle keltaisella taustavärillä. Näin organisaatio voi visualisoida maturiteettimallin kautta datanhallinnan nykytilan ja AI-valmiuden vaatiman datanhallinnan tavoitetilan eroa. Lisäksi ensimmäiseen sarakkeeseen on lisätty aineiston perusteella laadittu tekoälykehityksen näkökulmasta suositeltu datanhallinnan kehitystyön prioriteetti-järjestys tekoälykehitysvaiheittain.

Suositeltu kehitysprioriteetti (1=tukee ideoinnista lähtien, 2=tukee AI-projektin määrityksestä lähtien, 3=tukee datan kuratoinnista lähtien, 4=tukee prototyypin luonnista lähtien) ↓	datanhallinnan maturiteettiasteikko → datanhallinnan osa-alueet ↓	**1 (alustava / tapauskohtainen)** "Organisaation toiminnoissa ei sovelleta datanhallinnan parhaita käytäntöjä. Soveltuvia työkaluja ei ole saatavilla tai niitä ei käytetä."	**2 (toistettavissa / reaktiivinen)** "Osa organisaation liiketoiminta-alueista ja/tai funktioista käyttää suositeltuja prosesseja ja työkaluja, osa ei."	**3 (ennakoiva / määritelty)** "Organisaatiolla on dokumentoidut standardit johdonmukaiseen toimintaan ja soveltuvien työkalujen tehokkaaseen käyttöön."	**4 (hallittu)** "Olemassa olevat datanhallinnan prosessit, joita monitoroidaan. Suositellut työkalut ovat käytössä ja niitä käytetään johdonmukaisesti läpi organisaation."	**5 (optimoitu / tehokas)** "Sisäänrakennettuun toimintaan kohdistetaan uudelleenarviointia sekä toimintaa kehitetään ja seurataan jatkuvasti."
1	**Metadatan hallinta**	Vähän tai ei ollenkaan kuvauksia datasta (data katalogit, data standardit) Ei metadatan hallinnan työkaluja	Kasvava tietoisuus datavarannoista Kehittyvät kuvaukset datasta (data katalogit, data standardit) Kehittyvä työkaluvalikoima	Skaalattavat prosessit ja työkalut Yhdenmukaiset datakuvaukset	Standardoidut prosessit ja työkalut Yhdenmukaiset datakuvaukset käytössä läpi organisaation	Jatkuva kehitys
1	**Datan hallinnointi**	Vähäistä tai olematonta Rajallinen työkaluvalikoima Siilokohtaisesti määritettyjä rooleja Ei jalkautettuja säännöstöjä Puutteellinen riskienhallinta	Kasvava tietoisuus datan merkityksestä Kehittyvä datan hallinnointi Ensimmäiset askeleet kohti yhtenäisiä työkaluja Joitain rooleja, vastuita ja prosesseja määritelty	Data nähdään liiketoiminnallisena mahdollistajana Koordinoitu säännöstöjen määrittely ja hallinta Skaalattavat prosessit ja työkalut Aiempaa vähemmän manuaalisia vaiheita Dataprosessit ovat aiempaa ennustettavampia	Keskitetty datan hallinnointi Datanhallinnan metriikat käytössä	Dataohjautuva kulttuuri Ennustettavat dataprosessit Vähentynyt riskitaso
1	**Data-arkkitehtuuri**	Rajallinen työkaluvalikoima	Kehittyvä työkaluvalikoima Kehittyvät data-arkkitehtuurikuvaukset datasta (tietomallit, tietovirrat)	Skaalattavat prosessit ja työkalut Yhdenmukaiset data-arkkitehtuurikuvaukset	Standardoidut prosessit ja työkalut Yhdenmukaiset data-arkkitehtuurikuvaukset käytössä läpi organisaation	Jatkuva kehitys
1	**Datan mallinnus ja suunnittelu**	Datan mallinnus ja suunnittelu on vähäistä tai olematonta Rajallinen työkaluvalikoima	Lähestymistapa datan mallinnukseen ja suunnitteluun määritelty Kehittyvä työkaluvalikoima Kehittyvät kuvaukset (datasanasto, määritelmät)	Skaalattavat työkalut Yhdenmukainen tapa kerätä liiketoimintavaatimuksia ja piirtää kuvauksia	Yhdenmukaiset työkalut ja tapa mallintaa ja suunnitella jalkautettu läpi organisaation	Jatkuva datan mallinnuksen ja suunnittelun käytäntöjen kehitys
2	**Tietoturva**	Rajallinen työkaluvalikoima Puutteellinen tietoturva ja/tai tietosuoja aiheuttaa yleisesti ongelmia	Kehittyvä työkaluvalikoima Joitain rooleja ja prosesseja määritelty Kasvava tietoisuus tietoturvan, tietosuojan ja riskienhallinnan merkityksestä Kehittyvät kuvaukset (datan luokittelu, tietoryhmät, pääsyrajoitukset...)	Skaalattavat prosessit ja työkalut Koordinoitu säännöstöjen määrittely ja hallinta Joitain yhdenmukaisia kontrollipisteitä	Standardoidut prosessit ja työkalut Organisaation laajuinen kyvykkyys auditointeihin	Jatkuva kehitys

Kuva 11. Tekoälykehityksen näkökulmasta painotettu datanhallinnan maturiteettimalli.

5.2 Kehittämistehtävän arviointi

Opinnäytetyön alussa taustoitettiin opinnäytetyön kehittämistehtävää sen relevanssin arvioimiseksi ja sen suhteen, miten eri organisaatiot voisivat opinnäytetyön tuloksista hyötyä. Kehittämistehtävän onnistumista voidaan mitata arvioimalla tuotoksen eli tekoälykehityksen mukaan painotetun datanhallinnan maturiteettimallin hyödynnettävyyttä muun muassa organisaatioiden AI:n hallinnointi -hankkeiden nykytila-analyysivaiheessa. Painotetun maturiteettimallin arviointia on mahdollista tehdä opinnäytetyön puitteissa vain AIGA-hankkeen jäsenten kesken siltä osin, miten nämä asiantuntijat kokevat, että opinnäytetyön tavoitteeseen on päästy ja miltä osin tuotokset suhtautuvat koko AIGA-hankkeen AI governance -osioon. Kehittämistehtävää voidaan arvioida myös sen mukaan, onnistuiko opinnäytetyö vastaamaan siinä asetettuihin tutkimuskysymyksiin sekä niihin liittyen nostamaan tekoälykehitykseen liittyvästä datanhallinnasta jotain uutta esille tai toisaalta vahvistamaan teoriapohjan datanhallinnan mittareiden pätevyyttä arvioitaessa organisaatioiden datanhallinnan kyvykkyyttä tekoälyn näkökulmasta. Lisäksi opinnäytetyössä pyrittiin asettamaan käsiteltyihin teemoihin perustuvia jatkotutkimusehdotuksia. Opinnäytetyön tehokkuutta voidaan arvioida myös käytettyjen resurssien eli ajan ja henkilöresurssien suhteen.

Kehittämistehtävään valittujen menetelmien validiteettia voidaan arvioida sen suhteen, ovatko menetelmät päteviä kyseessä olevan tavoitteen saavuttamiseen. Kehittämistehtävän reliabiliteettia arvioidaan taas siitä näkökulmasta, kuinka toistettavissa työ on. Kehittämistehtävän haastatteluihin ja ideointityöpajaan pyydettiin Loihde-konsernin konsulteista kokeneita datanhallinnan ja tekoälyn asiantuntijoita. Haastateltavat rajattiin pääasiassa Loihde Advisory Oy:n henkilökuntaan, koska painotettu datanhallinnan maturiteettimalli tulee heidän työkaluvalikoimaansa ja tuotoksen täytyy vastata heidän tarpeitansa datanhallinnan asiakashankkeissa. Haastattelupyyntöön liitettiin kehittämistyön tarkoitus ja mahdollisuus vastata haastatteluun luottamuksellisesti. Haastateltavilta kysyttiin erikseen lupa haastattelujen nauhoittamiseen sekä nimen ja tittelin julkaisuun opinnäytetyön yhteydessä. Kaikki haastateltavat antoivat luvan kaikkiin kohtiin. Koko kehittämistehtävän ajan opinnäytetyön kirjoittaja osallistui viikoittaisiin AIGA-tiimin palavereihin. Näiden palaverien kautta oli sekä mahdollisuus raportoida kehittämistyön etenemisestä että saada säännöllistä palautetta.

Opinnäytetyön kehittämistehtävän lopputuloksesta raportoidaan AIGA-tiimille ja tuotoksen validoinnin jälkeen työkalu viedään Loihde Advisory Oy:ssa kehitettyyn konsulttien yhteiseen työkalupakkiin sekä tästä viestitään koko henkilöstölle. Varsinaista opinnäytetyön tuotosta voidaan arvioida vasta opinnäytetyön jälkeen, kun painotettua

maturiteettimallia hyödynnetään asiakashankkeissa osana konsultin työkalupakkia. Työkalutiedoston latausmäärät Loihde Advisory Oy:n työkalupakista kertovat osaltaan sitä, kuinka paljon kyseistä työkalua hyödynnetään. Konsulteille kohdennettujen kyselyiden kautta voidaan myöhemmin arvioida tarkemmin, kuinka kattavasti työkalulla saadaan kartoitettua tarvittavat asiat ja nopeuttaako se nykytila-analyysien läpivientiä verrattuna tilanteeseen, jossa maturiteettimalli luotaisiin painotuksineen aivan alusta erikseen jokaista hanketta varten. Näiden arviointien perusteella saadaan kokonaisvaltaisempi näkemys siitä, tapahtuiko varsinaista muutosta. Lisäksi voidaan kerätä asiakasorganisaatioilta kommentteja siitä, miten hyödyllisenä he kokivat käytetyn painotetun datanhallinnan maturiteettimallin.

Koska opinnäytetyön kirjoittaja tekee opinnäytetyön työnantajalleen, on syytä pohtia myös kirjoittajan puolueellisuutta suhteessa kehittämistehtävään. Työssä pyrittiin puolueettomuuteen pysymällä neutraalina suhteessa haastateltaviin kollegoihin esimerkiksi pysyttelemällä haastatteluissa ja ideointityöpajassa taustalla vastaamisen ja ideointivaiheen aikana, jotta vastaukset olisivat yksilön mielipiteitä ja työpajassa saatu tulos olisi ryhmän tulos.

5.3 Tavoitteiden saavuttamisen ja tulosten arviointi

Tämän opinnäytetyön tarkoituksena oli tarkastella datanhallinnan merkitystä tekoälykehityksessä. Tarkoitus auttoi rajaamaan työtä datanhallinnan osa-alueisiin ja tekoälykehityksen vaiheisiin. Opinnäytetyön tavoitteena oli luoda tekoälykehityksen mukaan painotettu datanhallinnan maturiteettimalli, jolla voidaan kartoittaa organisaatioiden datanhallinnan eri osa-alueiden kyvykkyyttä valjastaa liiketoimintadataa tekoälyn käyttöön. Tavoitteen kautta työhön sisällytettiin lisäksi eri datanhallinnan maturiteetti- ja AI-hallinnointimallien tarkastelu.

Maturiteettimallin kohdekäyttäjäryhmänä ovat Loihde-konsernin datanhallinnan konsultit, joiden asiakashankkeisiin sisältyy monesti datanhallinnan nykytilan ja tavoitetilan selvitystyö. Painotettu datanhallinnan maturiteettimalli rakennettiin sekä teoriataustaa että asiantuntijahaastatteluista ja ideointityöpajasta kerättyjä havaintoja vasten. Opinnäytetyön tuotoksen kohdekäyttäjäryhmän perusteella haastatteluihin valittiin sekä datanhallinnan että tekoälyn pitkän linjan asiantuntijoita, jotta näiden alueiden kokemuskenttä olisi mahdollisimman laajasti edustettuna. Koska asiantuntijat kuuluvat kohdekäyttäjäryhmään, oli lisäksi perusteltua kuulla heidän arvokkaita havaintojaan datanhallinnan merkityksestä tekoälykehityksessä. Yksitoista haastattelua oli riittävä määrä saturaation täyttymiseen.

Asiantuntijoiden työkokemus kattoi lopulta kaikki datanhallinnan osa-alueet ja tekoälykehityksen vaiheet.

Jokaisessa haastattelussa hyödynnettiin ennalta laadittua teoriaan perustuvaa haastattelurunkoa, jolloin myös kerättyjä vastauksia oli helpompi vertailla keskenään. Kaikki haastatteluvastaukset validoitiin lähettämällä litteroitu aineisto kullekin haastateltavalle tarkistettavaksi sähköpostitse. Kaikki vastaukset validoitiin ja vain muutamassa tapauksessa vastauksiin ehdotettiin pieniä muutoksia, jotka nekin olivat lähinnä kirjoitusvirheitä. Haastatteluista ja ideointityöpajasta kerätty aineisto analysoitiin siten, että jokaista havaintoa kohden esitettiin havaintoon liittyvä katkelma haastatteluvastauksista, jolloin myös lukija voi nähdä perusteet johtopäätöksille.

Opinnäytetyön etenemisestä ja lopullisen tuotoksen rakentamisesta raportoitiin viikoittaisissa Loihde Advisoryn sisäisissä AIGA-hankkeen palavereissa koko kehittämistehtävän ajan. Kehittämistehtävän lopuksi painotettua datanhallinnan maturiteettimallia verrattiin yhden johtavan AI-valtion, Singaporen luoman AI-hallintamallin sisältöön datanhallinnan aihealueiden osalta. Näillä kaikilla mainituilla menetelmillä varmistettiin, että lopullinen tuotos on riittävän kattava ja validi työkalu datanhallinnan maturiteetin analysointiin tekoälykehitykseen lähtevissä organisaatioissa. Sekä teoriataustassa että kerätyssä aineistossa korostuu datanhallinnan tärkeys dataa hyödyntävissä vaativissa hankkeissa, kuten tekoälykehityksessä. Voidaan siis sanoa, että tutkimus on toistettavissa samanlaisine havaintoineen.

6 Johtopäätökset

Tekoälykehityksen onnistuminen on vahvasti riippuvainen riittävästä datanhallinnan maturiteetista. Tekoälykehitykseen lähdettäessä tarvitaan ennakoivaa datanhallinnan maturiteettia kaikkien muiden paitsi dokumenttien ja sisällönhallinnan osalta, jotta perusteet tuotantokelpoiselle tekoälyratkaisulle on rakennettu jo alusta lähtien. Jotta lisäksi varmistetaan, ettei jäädä tekoälyratkaisujen kokeiluvaiheeseen, vaan edetään kohti tuotantokelpoista ja skaalautuvaa AI-valmiutta, tarvitaan lisäksi hallittua maturiteettia sekä datan laadun, datavarastot ja analytiikan että datan hallinnoinnin kyvykkyyksien osalta. AI-valmis organisaatio ylläpitää liiketoimintastrategian ja datan välistä liittoa hallinnoinnin kautta, jotta saadaan tuotettua arvoa tuottavia tekoälyratkaisuja.

Opinnäytetyön tuloksia voidaan hyödyntää arvioinnin tukena, kun organisaatiot haluavat selvittää datanhallinnan kyvykkyyttä tekoälykehitystä ajatellen. Lisäksi tuloksia voidaan hyödyntää sekä asettamaan datanhallinnan tavoitematuriteetti sille tasolle, joka palvelee parhaiten tuotantokelpoisen ja aidosti liiketoimintahyödyllisen tekoälyratkaisun kehittämistä, että määrittämään datanhallinnan kehitysaskeleet, jotka parhaiten tukevat tekoälykehitystä.

Opinnäytetyössä keskityttiin datanhallinnan ja tekoälyn asiantuntijoiden havaintoihin datanhallinnan merkityksestä tekoälykehityksessä. Jatkotutkimuksena olisi kiinnostavaa tietää, miten eri organisaatiot ovat vieneet tekoälyhankkeita läpi datanhallinnan eri osa-alueiden näkökulmasta ja miten sekä alhainen että korkeampi datanhallinnan maturiteetti on vaikuttanut sekä hankkeisiin että laajemmin organisaation liiketoimintaan ja kilpailukykyyn. Lisäksi olisi mielenkiintoista tutkia, millä muilla kuin datanhallinnan osa-alueilla on vaikutusta arvioitaessa organisaation AI-valmiutta.

Opinnäytetyöaiheen ajankohtaisuus kävi ilmi läpi kehittämishankkeen. Uusia tekoälyn hallinnointia käsitteleviä julkaisuja oli usein ja säännöllisesti saatavilla. Varmasti merkittävimpänä näistä voidaan pitää huhtikuussa 2021 Euroopan komission julkaisemaa asetusehdotusta tekoälyn harmonisoidusta sääntelystä. Tämä tuo painetta kaikille organisaatioille täyttää tulevat vaatimukset tekoälyn hyödyntämiseksi. Iso osa vaatimuksista sisältyy datan hallinnoinnin ja datanhallinnan piiriin, joiden osalta useat organisaatiot ovat pahasti perässä. Esimerkiksi iso osa Suomen suurimmista organisaatioista on investoimassa tekoälyyn lähivuosien aikana, mutta samalla suurin osa näistä organisaatioista kokee, että he ovat sekä hankkeissaan että datan hallintaratkaisujen osalta vielä melko tai täysin alkutekijöissään (Professio 2021). On selvää, että kilpailu AI-valmiuden herruudesta käydään datanhallinnan kentällä.

Lähteet

Ahopelto, M. 2019. Data Governance 3.0. Luettavissa: https://www.rootsof.ai/blog/data-governance. Luettu: 23.5.2021.

Anderson, J. & Coveyduc, J. 2020. Artificial intelligence for business: a roadmap for getting started with AI. John Wiley & Sons, Inc. Hoboken, New Yersey.

Combs, V. 2021. Gartner: AI is moving fast and will be ready for prime time sooner than you think. Luettavissa: https://www.techrepublic.com/article/gartner-ai-is-moving-fast-and-will-be-ready-for-prime-time-sooner-than-you-think/. Luettu: 16.9.2021.

European Commission. 2021. Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Luettavissa: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206. Luettu: 20.10.2021.

Gupta, N. & Mangla, R. 2020. Artificial Intelligence Basics: A Self-Teaching Introduction. Mercury Learning & Information. Herndon, Virginia.

Ilveskero, N. 2021. Miten rakennetaan AI-valmis organisaatio ja miksi pian on pakko? Luettavissa: https://ai-governance.eu/miten-rakennetaan-ai-valmis-organisaatio/. Luettu: 17.5.2021.

ISO/IEC DIS 38507. Information technology. Governance of IT. Governance implications of the use of artificial intelligence by organizations. 2021. Luettavissa: https://www.iso.org/obp/ui/#iso:std:iso-iec:38507:dis:ed-1:v1:en. Luettu: 20.10.2021.

ISO/IEC TR 38505. Information technology. Governance of IT. Governance of data. Part 2: Implications of ISO/IEC 38505-1 for data management. 2018. Luettavissa: https://www.iso.org/obp/ui/#iso:std:iso-iec:tr:38505:-2:ed-1:v1:en. Luettu: 19.10.2021.

ISO/TS 8000-60. Data quality. Part 60: Data quality management: Overview. 2017. Luettavissa: https://www.iso.org/obp/ui/#iso:std:iso:ts:8000:-60:ed-1:v1:en. Luettu: 20.10.2021.

IT Governance Privacy Team, I. T. G. 2020. EU General Data Protection Regulation (GDPR) An implementation and compliance guide, fourth edition. IT Governance Publishing. Ely.

Ojasalo, K., Moilanen, T. & Ritalahti, J. 2015. Kehittämistyön menetelmät – Uudenlaista osaamista liiketoimintaan. Sanoma Pro Oy. Helsinki.

PDPC. 2020. Model Artificial Intelligence Governance Framework. Luettavissa: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf. Luettu: 29.10.2021.

Professio. 2021. Tutkimus tietohallintojohdon kehityshankkeista 2022–2024 – Päättäjäkysely: Strategy Talk CIO 2022 -osallistujat. Luettavissa: https://professio.fi/paattajaraportti-strategy-talk-cio/. Luettu: 3.11.2021.

Roe, C. 2011. Assessing Data Management Maturity Using the DAMA DMBOK Framework. Luettavissa: https://www.dataversity.net/assessing-data-management-maturity-using-the-dama-dmbok-framework-%E2%80%93-part-1/#. Luettu: 25.10.2021.

Thomas, J. 2019. Operationalizing AI — Managing the End-to-End Lifecycle of AI. Luettavissa: https://medium.com/inside-machine-learning/ai-ops-managing-the-end-to-end-lifecycle-of-ai-3606a59591b0. Luettu: 23.5.2021.

Sebastian-Coleman, L. 2018. Navigating the Labyrinth – An Executive Guide to Data Management. Technics Publications. New Jersey.

Sebastian-Coleman, L. 2020. CDMP Study Group. Session 7 – Data Management Capability Maturity Models. Luettavissa: https://damanewengland.org/images/downloads/CDMP_Study_Group/data_management_capability_maturity_models.pdf. Luettu: 25.10.2021.

Taylor, K. 2020. Data governance maturity models explained. Luettavissa: https://www.hitechnectar.com/blogs/data-governance-maturity-models-explained/. Luettu: 23.5.2021.

Technics Publications. 2017. DAMA Guide to the Data Management Body of Knowledge. Basking Ridge, New Jersey.

Turun Yliopisto. 2020. Turun yliopiston johtamalle konsortiolle miljoonarahoitus tekoälyn hallintamallien tutkimukseen. Luettavissa: https://www.utu.fi/fi/ajankohtaista/mediatiedote/turun-yliopiston-johtamalle-konsortiolle-miljoonarahoitus-tekoalyn. Luettu: 2.11.2021.

Liitteet

Liite 1. Haastattelukysymykset

Yleistä

1. Millä datanhallinnan osa-alueilla olet työskennellyt?
2. Missä organisaatioiden tekoälykehityksen vaiheissa olet ollut mukana tai sinulla on ollut näkyvyys niihin?
3. Mitkä datanhallinnan ja/tai tekoälyn standardit ja regulaatiot ovat tutuimpia työssäsi?
4. Mitä datanhallinnan ja/tai tekoälyn maturiteettimalleja olet hyödyntänyt työssäsi?

Hyvä datanhallinta

5. Miten hyvä datanhallinta ilmenee?
6. Mikä merkitys hyvällä datanhallinnalla on organisaatioissa, jotka haluavat hyödyntää tekoälyä? Mitä konkreettisia lyhyen- ja pitkän aikavälin hyötyjä hyvästä datanhallinnasta on näille organisaatioille?
7. Mitä tulisi huomioida niiden organisaatioiden datanhallinnan maturiteetin arvioinnissa, jotka harkitsevat tekoälyn hyödyntämistä tai jo hyödyntävät sitä?

Datanhallinnan rooli tekoälykehityksessä

8. Käydään läpi tekoälykehitys vaihe vaiheelta:
    a. Mitkä datanhallinnan osa-alueet ovat kriittisiä tekoälykehityksen vaiheessa X (ideointi, AI-projektin määritys, datan kuratointi, prototyypin luonti, tuontantoon siirtäminen, AI:n elinkaaren hallinta) ja miksi?
    b. Millainen datanhallinnan maturiteettitaso vaaditaan AI-valmiille organisaatiolle vaiheessa X (ideointi, AI-projektin määritys, datan kuratointi, prototyypin luonti, tuontantoon siirtäminen, AI:n elinkaaren hallinta) asteikolla 1-5 (1 = alustava tai tapauskohtainen, 2 = reaktiivinen, 3 = ennakoiva ja määritelty, 4 = hallittu, 5 = optimoitu ja tehokas)?

Liite 2. Haastatteluaineisto

Bergman, T. 2021. Toimitusjohtaja. Loihde Advisory Oy. Haastattelu 06.08.2021.

Ilveskero, N. 2021. Myynti- ja markkinointijohtaja. Loihde Advisory Oy. Haastattelu 11.08.2021.

Kangas-Lång, K. 2021. Palvelujohtaja & Johtava konsultti, Datan hallinta ja laatu. Loihde Advisory Oy. Haastattelu 06.08.2021.

Keränen, L. 2021. Palvelujohtaja, Data-alustat ja raportointi. Loihde Advisory Oy. Haastattelu 09.08.2021.

Laatikainen, T. 2021. Johtaja, Data & Analytiikka -palvelut. Loihde Advisory Oy. Haastattelu 13.08.2021.

Lahtinen, T. 2021. Johtava Analytiikka- ja tekoälyjohtaja. Loihde Analytics Oy. Haastattelu 12.08.2021.

Masala, S. 2021. Johtava neuvonantaja. Loihde Advisory Oy. Haastattelu 17.08.2021.

Nikkilä, T. 2021. Koneoppimisinsinööri. Loihde Factor Oy. Haastattelu 10.08.2021.

Peltomäki, J. 2021. Konsultti, Data-alustat ja raportointi. Loihde Advisory Oy. Haastattelu 09.08.2021.

Rönkä, S. 2021. Koneoppimisinsinööri. Loihde Factor Oy. Haastattelu 13.08.2021.

Vartiainen, S. 2021. Johtava konsultti & Palvelualueen vetäjä, Tietosuoja. Loihde Advisory Oy. Haastattelu 06.08.2021.

Liite 3. Tuotos: tekoälykehityksen mukaan painotettu datanhallinnan maturiteetti-malli

Suositeltu kehitysprioriteetti (1=tukee ideoinnista lähtien, 2=tukee AI-projektin määrityksestä lähtien, 3=tukee datan kuratoinnista lähtien, 4=tukee prototyypin luonnista lähtien) ↓	**datanhallinnan maturiteettiasteikko** → **datanhallinnan osa-alueet** ↓	**1 (alustava / tapauskohtainen)** "Organisaation toiminnoissa ei sovelleta datanhallinnan parhaita käytäntöjä. Soveltuvia työkaluja ei ole saatavilla tai niitä ei käytetä."	**2 (toistettavissa / reaktiivinen)** "Osa organisaation liiketoiminta-alueista ja/tai funktioista käyttää suositeltuja prosesseja ja työkaluja, osa ei."	**3 (ennakoiva / määritelty)** "Organisaatiolla on dokumentoidut standardit johdonmukaiseen toimintaan ja soveltuvien työkalujen tehokkaaseen käyttöön."	**4 (hallittu)** "Olemassa olevat datanhallinnan prosessit, joita monitoroidaan. Suositellut työkalut ovat käytössä ja niitä käytetään johdonmukaisesti läpi organisaation."	**5 (optimoitu / tehokas)** "Sisäänrakennettuun toimintaan kohdistetaan uudelleenarviointia sekä toimintaa kehitetään ja seurataan jatkuvasti."
1	**Metadatan hallinta**	Vähän tai ei ollenkaan kuvauksia datasta (data katalogit, data standardit) Ei metadatan hallinnan työkaluja	Kasvava tietoisuus datavarannoista Kehittyvät kuvaukset datasta (data katalogit, data standardit) Kehittyvä työkaluvalikoima	Skaalattavat prosessit ja työkalut Yhdenmukaiset datakuvaukset	Standardoidut prosessit ja työkalut Yhdenmukaiset datakuvaukset käytössä läpi organisaation	Jatkuva kehitys
1	**Datan hallinnointi**	Vähäistä tai olematonta Rajallinen työkaluvalikoima Siilokohtaisesti määritettyjä rooleja Ei jalkautettuja säännöstöjä Puutteellinen riskienhallinta	Kasvava tietoisuus datan merkityksestä Kehittyvä datan hallinnointi Ensimmäiset askeleet kohti yhtenäisiä työkaluja Joitain rooleja, vastuita ja prosesseja määritelty	Data nähdään liiketoiminnallisena mahdollistajana Koordinoitu säännöstöjen määrittely ja hallinta Skaalattavat prosessit ja työkalut Aiempaa vähemmän manuaalisia vaiheita Dataprosessit ovat aiempaa ennustettavampia	Keskitetty datan hallinnointi Datanhallinnan metriikat käytössä	Dataohjautuva kulttuuri Ennustettavat dataprosessit Vähentynyt riskitaso
1	**Data-arkkitehtuuri**	Rajallinen työkaluvalikoima	Kehittyvä työkaluvalikoima Kehittyvät data-arkkitehtuurikuvaukset datasta (tietomallit, tietovirrat)	Skaalattavat prosessit ja työkalut Yhdenmukaiset data-arkkitehtuurikuvaukset	Standardoidut prosessit ja työkalut Yhdenmukaiset data-arkkitehtuurikuvaukset käytössä läpi organisaation	Jatkuva kehitys
1	**Datan mallinnus ja suunnittelu**	Datan mallinnus ja suunnittelu on vähäistä tai olematonta Rajallinen työkaluvalikoima	Lähestymistapa datan mallinnukseen ja suunnitteluun määritelty Kehittyvä työkaluvalikoima Kehittyvät kuvaukset (datasanasto, määritelmät)	Skaalattavat työkalut Yhdenmukainen tapa kerätä liiketoimintavaatimuksia ja piirtää kuvauksia	Yhdenmukaiset työkalut ja tapa mallintaa ja suunnitella jalkautettu läpi organisaation	Jatkuva datan mallinnuksen ja suunnittelun käytäntöjen kehitys
2	**Tietoturva**	Rajallinen työkaluvalikoima Puutteellinen tietoturva ja/tai tietosuoja aiheuttaa yleisesti ongelmia	Kehittyvä työkaluvalikoima Joitain rooleja ja prosesseja määritelty Kasvava tietoisuus tietoturvan, tietosuojan ja riskienhallinnan merkityksestä Kehittyvät kuvaukset (datan luokittelu, tietoryhmät, pääsyrajoitukset...)	Skaalattavat prosessit ja työkalut Koordinoitu säännöstöjen määrittely ja hallinta Joitain yhdenmukaisia kontrollipisteitä	Standardoidut prosessit ja työkalut Organisaation laajuinen kyvykkyys auditointeihin	Jatkuva kehitys
2	**Datan integrointi ja yhteentoimivuus**	Rajallinen integrointikyvykkyys Data on huonosti yhteentoimivaa	Kasvava tietoisuus datan yhteentoimivuuden merkityksestä Kehittyvät integrointikyvykkyydet	Kehittyvä datan yhteentoimivuus	Mitattava parannus datan yhteentoimivuudessa	Jatkuva kehitys
2	**Datan laatu**	Rajallinen työkaluvalikoima Datan laatuongelmat (sisältäen datan kattavuus- ja saatavuusongelmat) ovat yleisiä, mutta niitä ei käsitellä	Kasvava tietoisuus datan laatuongelmien (sisältäen datan kattavuus- ja saatavuusongelmien) merkityksestä Kehittyvä työkaluvalikoima Joitain rooleja ja prosesseja määritelty	Skaalattavat prosessit ja työkalut Joitain kontrollipisteitä Kehittyvä datan laatu (sisältäen datan kattavuuden ja saatavuuden)	Mitattavaa kehitystä datan laadussa (sisältäen datan kattavuuden ja saatavuuden) Dataan liittyvien riskien hallinta	Datan jakamista valvotaan turhien duplikaattien estämiseksi Hyvin ymmärretyt metriikat datan laadun (sisältäen datan kattavuuden ja saatavuuden) ja dataprosessien laadun hallinnoimiseksi
2	**Datan tallennus ja toiminnot**	Vähän tai ei ollenkaan datan elinkaarikuvauksia	Joitain rooleja ja dataprosesseja määritelty Kehittyvät datan elinkaarikuvaukset	Vähemmän manuaalista puuttumista dataprosesseihin. Skaalattavat prosessit	Standardoidut prosessit ja työkalut Datan elinkaarta monitoroidaan	Jatkuva kehitys
2	**Datavarastot ja analytiikka**	Toiminnot muutamien asiantuntijoiden varassa	Kehittyvät data-alustat ja analytiikka Määritellyt roolit Joitain rooleja ja prosesseja määritelty	Tulokset paremmin ennakoitavissa	Standardoidut prosessit ja työkalut Organisaationlaajuinen kyvykkyys	Prosessiautomaatio Jatkuva kehitys Näkyvyys dataan läpi prosessien
2	**Viite- ja ydintiedon hallinta**	Rajalliset työkalumahdollisuudet hallita ko. dataa	Kehittyvä työkaluvalikoima Kasvava ymmärrys ko. konsepteista	Datan hyödynnettävyys läpi organisaation	Standardoidut prosessit ja työkalut	Jatkuva kehitys
3-4* *2, Jos tekoälykehityksessä hyödynnetään strukturoimatonta dataa	**Dokumenttien ja sisällönhallinta**	Rajalliset työkalumahdollisuudet hallita ko. dataa Tekoälymallien dokumentointi on vähäistä tai olematonta	Kehittyvä työkaluvalikoima Kasvava ymmärrys ko. konsepteista Kehittyvät kuvaukset tekoälymalleista	Datan hyödynnettävyys läpi organisaation Yhdenmukaiset kuvaukset tekoälymalleista.	Standardoidut prosessit ja työkalut Yhdenmukaiset, koneellisesti luotavat kuvaukset tekoälymalleista käytössä läpi organisaation	Jatkuva kehitys