Tatu Niskanen

MASSADATAN HYÖDYNTÄMINEN TIETOTURVALLISUUDEN NÄKÖKULMASTA

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2021

TIIVISTELMÄ

Niskanen, Tatu
Massadatan hyödyntäminen tietoturvallisuuden näkökulmasta
Jyväskylä: Jyväskylän yliopisto, 2021, 33 s.
Tietojärjestelmätiede, Kandidaatin tutkielma
Ohjaaja(t): Kyppö, Jorma

Massadata, johon viitataan yleisesti myös käsitteellä Big data, on kokoelma dataa, joka on kooltaan todella suuri, se on luonteeltaan monipuolista ja hajanaista, ja sitä tulee nopeasti lisää. Massadata aiheuttaa organisaatioille hyötyjen ja mahdollisuuksien lisäksi myös monia tietoturvariskejä. Digitalisoitumisen mukanaan tuomat työkalut, kuten massadata ja sen useat sovellukset, halutaan usein nähdä vain taloudellisen kasvun ja tehokkuuden mahdollisuuksien kautta. Niiden todellista kokonaisvaltaisuutta ei aina hahmoteta riittävän kattavasti. Siksi tässä tutkielmassa tarkastellaan massadatan hyödyntämistä organisaatioissa sen aiheuttamien tietoturvariskien näkökulmasta. Ensin tutkielmassa selvennetään, miten eri organisaatioissa kerätään, käsitellään ja hyödynnetään massadataa, ja mihin sitä soveltamalla tyypillisesti pyritään. Massadatan hyödyntämisessä on kyse massiivisen raakadatamäärän keräämisestä, jota säilötään tyypillisesti usealla eri palvelimella tai muulla tallennusmedialla. Tästä raakadatasta pyritään jalostamaan hyödyllistä informaatiota data-analytiikan menetelmillä.

Tämän jälkeen tutkielmassa käydään läpi, millaisia tietoturvariskejä massadatan hyödyntämisestä aiheutuu. Massadata aiheuttaa tietoturvariskejä erityisesti GDPR-aikana korostuneen henkilötietosuojan näkökulmasta, mutta siihen liittyy myös muita vähemmän itsestään selviä tietoturvahaasteita. Nämä liittyvät esimerkiksi massadatan varastointiin, haitallisen toiminnan tunnistamiseen suurien datamassojen seasta, sekä salausalgoritmien suorituskykyyn. Tutkielmassa tarkastellaan myös tiiviisti, miten organisaatioissa voidaan hallita joitakin keskeisimmistä massadatan tietoturvariskeistä. Tutkielma on toteutettu kokonaan kirjallisuuskatsauksena.

Asiasanat: massadata, big data, data, datanhallinta, tietoturva, kyberturvallisuus, organisaatiot, tietoturvariski, riskienhallinta, tietosuoja, GDPR, tietovuodot, salaus

ABSTRACT

Niskanen, Tatu
Big data utilization and related information security risks
Jyväskylä: University of Jyväskylä, 2021, 33 pp.
Information Systems Science, Bachelor's thesis
Supervisor(s): Kyppö, Jorma

Big data is a set of data that is large in size, diverse by nature, and rapidly increasing. In addition to the benefits, big data causes notable information security risks to organizations. The tools brought by digitalization are often seen only trough their benefits and their security issues are not fully taken into concern. This dissertation examines the utilization of big data and the related information security risks and problems it arises for organizations. First, it is clarified to the reader how big data is collected, processed, and utilized in different organizations, and for what purpose. This is followed by a review of the information security risks involved. After that we will have a look at how organizations can manage some of these risks. Finally, the summary provides an assessment and conclusion about the usage of big data, and the related information security challenges. We also summarize briefly what organizations should consider when leveraging big data in their operations. The dissertation is carried out as a literature review.

Keywords: big data, data, information security, cybersecurity, organizations, information security risk, risk management, data protection, data privacy, GDPR, data leak, encryption

KUVIOT

SISÄLLYS

1 Johdanto

Datan määrä kasvaa maailmassa jatkuvasti digitalisaation myötä. Vielä vuonna 2000, vain noin neljännes maailman tietovarannoista oli varastoitu digitaalisena datana, kun vuonna 2013 tämä luku oli jo yli 98 % (Mayer-Schönberger ja Cukier, 2013). Tämän lisäksi esimerkiksi vuonna 2017 arvioitiin, että kaikesta maailman datasta noin 80–91 % oli luotu edellisen kahden vuoden aikana. Väittämä, jonka mukaan 90 % kaikesta maailman datasta on luotu edellisen kahden vuoden aikana, on esiintynyt useassa tutkimuksessa ja artikkelissa eri vuosina, mutta laskennallisesti se vaikuttaa pitävän kutakuinkin paikkansa. (Nathan, Nicola, Roger & Kiersten, 2017.)

Massadata on käsitteenä helppo ymmärtää väärin, joten sen täsmällinen määrittely on tutkielman kannalta oleellista. Massadata, tunnettu yleisesti myös nimellä ”Big Data”, on kokoelma dataa, joka on kooltaan todella suuri, se on luonteeltaan monipuolista ja hajanaista, ja sitä tulee nopeasti lisää. Käsitteen määrittelemiseksi on esitetty useampia malleja, mutta yleisimmän hyväksyttynä käytetyn mallin mukaan nämä kolme attribuuttia määrittelevät massadatan. Tämä tunnetaan yleisesti ”kolmen V:n” mallina, englanninkielisten sanojen ”volume", ”velocity” ja ”variety” mukaisesti. (Demchenko, Gruengard & Klous, 2014.) Massadataksi voidaan kutsua siis mitä tahansa suurta kokoelmaa dataa, joka noudattaa näitä attribuutteja riittävällä tasolla. Näiden attribuuttien toteutumisen tasolla ei kuitenkaan ole alalla yleisesti hyväksyttyä kiinteää standardia, jonka mukaisesti datajoukko voidaan katsoa määrittyvän massadataksi. Massadatan määrittely jääkin tästä syystä ajoittain tulkinnanvaraiseksi. Tässä yhtenä mittapuuna käytetään vertailua ”tavanomaisiin” datajoukkoihin. Jos edellä mainitut vaatimukset täyttyvät ja tavanomaiset, vanhemmat datanhallinnan keinot ja työkalut, kuten perinteiset relaatiotietokannat eivät tämän vuoksi riitä, voidaan datajoukon katsoa olevan massadataa (Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011).

”Big data” on ollut terminä käytössä jo 1990-luvulta alkaen. Viime vuosina sitä ollaan alettu osittain käyttää enemmän datan edistyneiden analysointiratkaisujen ja -menetelmien kontekstissa, ja termin käyttäminen itse suurista datajoukoista puhuttaessa on jokseenkin vähentynyt. Käsitteellä siis viitataan nyky-

ään usein suurien ja monimutkaisten datajoukkojen käsittelyyn alana. Tämä selittyy sillä, että nykyään hyödynnettävän datan suuri koko on niin tyypillistä. (Boyd & Crawford, 2011.) Tässä tutkielmassa massadatan käsitteellä viitataan edellä esitettyjen ”kolmen V:n” mukaisiin suuriin datajoukkoihin, ja niiden käsittelyyn organisaatioiden toiminnassa.

Massadata on edelleen yksi tämän hetken puhutuimmista IT-alan trendeistä, ja sitä sovelletaan lukuisiin eri tarkoituksiin eri organisaatioissa. Massadataa hyödynnetään liki jokaisella alalla, ja sen hyödyntäminen on edelleen jatkuvassa nousussa. Esimerkiksi ETLAn vuonna 2015 toteuttaman kyselytutkimuksen mukaan, johon vastasi 1189 suomalaista yritystä, noin neljännes vastanneista yrityksistä hyödynsi massadataa toiminnassaan, ja sen hyödyntäminen oli selkeästi nouseva trendi yritysten keskuudessa. (Antikainen, Eskelinen, Koski, Niemi, Pajarinen, Pyykkönen ja de Vries, 2016.) Edeltävänä vuonna EU-alueen datatalouden arvoksi arvioitiin liki 250 miljardia euroa, joka kattaa 1,8 prosenttia alueen BKT:stä, ja maailman massadatateknologioiden ja -palveluiden markkinoille arvioitiin keskimäärin 23 prosentin vuosikasvua (IDC, 2015). Massadatan, ja siihen liittyvien haasteiden ymmärtäminen on siis oleellista myös tulevaisuutta ajatellen.

Massadatan hyödyntäminen aiheuttaa organisaatioille hyötyjen lisäksi myös monia tietoturvariskejä ja -ongelmia (N. Chaudhari ja S. Srivastava, 2016). Tässä tutkielmassa keskitytään tarkastelemaan ja määrittelemään näitä riskejä ja ongelmia organisaatioiden näkökulmasta. Alan keskeisten trendien tietoturvallisuuden kriittinen tarkastelu on tärkeää, sillä tekniikan hyödyntämisen lisääntyessä myös sen turvallisuuden varmistamisen merkitys kasvaa.

Maailman datavarannot ovat pääosin organisaatioiden hallitsemia, mutta suurin osa uudesta datasta syntyy kuitenkin yksityishenkilöiden tuottamana. Joidenkin arvioiden mukaan jopa 80 prosenttia kaikesta datasta on yritysten hallinnoimaa, ja siitä tuottavat 70 prosenttia yksityishenkilöt (Computer Sciences Corporation, 2014). Tämän takia yksi massadatan keskeinen haaste on ihmisten yksityisyydensuojan, eli tietosuojan säilyttäminen (Zhang, 2018). Kuinka organisaatio voi varmistua siitä, että heidän käsittelemänsä data ei sisällä henkilötietoja, eikä ajaudu vääriin käsiin? Vuonna 2018 voimaan astuneen GDPR-tietosuojasäädöksen myötä tämä on muodostunut erittäin keskeiseksi ongelmaksi massadatan hyödyntämisessä. Yksityisyydensuoja ei kuitenkaan ole ainoa massadatan aiheuttama tietoturvaongelma, vaan se esittää tämän lisäksi myös muita riskejä, jotka liittyvät esimerkiksi datan keräämiseen, varastointiin ja käsittelyyn (N. Chaudhari ja S. Srivastava, 2016). Tutkimuksessa tullaan perehtymään myös näihin kattavasti.

Aloitamme tutkielman tarkastelemalla, miten massadataa käytännössä hyödynnetään eri organisaatioissa, ja millaisia tavoitteita organisaatioilla on sen hyödyntämiselle. Käytämme tämän osion tilastollisessa osuudessa tarkastelussa kohdejoukkona pääasiassa suomalaisia yrityksiä, mutta vertaamme näitä myös ulkomaisiin toimijoihin kattavan kuvan takaamiseksi. Tätä kautta saadaan selvennettyä lukijalle massadatan käsitettä ja sen sovelluksia riittävän kattavasti tarkastellaksemme siihen liittyviä tietoturvaongelmia. Osion jälkeen siirrytään

tarkastelemaan tarkemmin, millaisia tietoturvariskejä ja -ongelmia organisaatioille aiheutuu massadatan hyödyntämisestä. Tutkielmassa tarkastellaan lopuksi, kuinka joihinkin massadatan aiheuttamiin tietoturvariskeihin voidaan organisaatioissa varautua.

Tutkielma avaa lukijalle massadatan käsitettä ja sen hyödyntämistä eri organisaatioissa tarjotakseen perustason ymmärryksen massadatasta ja sen sovelluksista, mutta ensisijaisesti sen tarkoitus on käsitellä tästä aiheutuvia tietoturvaongelmia. Tutkielman laajuuden rajaamiseksi pääpaino pidetään siis massadatan käsittelyssä sen aiheuttamien tietoturvariskien ja -ongelmien kontekstissa.

Tutkielman tarkoitus on vastata seuraaviin tutkimuskysymyksiin:

1. Miten organisaatiot hyödyntävät massadataa toiminnassaan?
2. Millaisia tietoturvaongelmia massadatan hyödyntäminen aiheuttaa?

Ensimmäiseen tutkimuskysymykseen vastataan luvussa 2; Massadatan hyödyntäminen ja sen tavoitteet organisaatioissa. Toiseen tutkimuskysymykseen vastataan luvussa 3; Massadatan tietoturvariskit ja niiden hallitseminen organisaatioissa. Luvussa kolme käsitellään erikseen massadatan tietosuojaan liittyviä tietoturvaongelmia ja muita tietoturvaongelmia yleisemmällä tasolla omissa alaluvuissaan. Lopuksi luvussa kolme käsitellään myös lyhyesti, kuinka joihinkin massadatan tietoturvaongelmiin voidaan organisaatioissa varautua.

2 Massadatan hyödyntäminen ja sen tavoitteet organisaatioissa

Tässä luvussa pohjustetaan tulevaa massadatan tietoturvaongelmien käsittelyä selventämällä, millä eri tavoilla organisaatiot keräävät ja hyödyntävät massadataa toiminnassaan, ja mihin tarkoituksiin. Luvun on tarkoitus vastata tutkielman ensimmäiseen tutkimuskysymykseen ”Miten organisaatiot hyödyntävät massadataa toiminnassaan?”. Ensimmäisessä alaluvussa selvennetään massadatan soveltamisen yleisimpiä tavoitteita ja käytänteitä organisaatioissa tilastojen avulla. Toisessa alaluvussa selvennetään tarkemmin, miten eri organisaatiot käytännössä keräävät, varastoivat ja käsittelevät massadataa hyödyntääkseen sitä.

2.1 Massadatan hyödyntäminen organisaatioissa tilastoilla havainnollistettuna

Osiossa valaistaan massadatan hyödyntämistä tilastojen avulla. Massadatan käytön tilastollinen havainnollistaminen toteutetaan pääasiassa Elinkeinoelämän Tutkimuslaitos ETLAn vuonna 2016 julkaiseman suomalaisen massadatakyselytutkimuksen, ”Massadatasta liiketoimintaa ja tehokkaita julkisia palveluja”, pohjalta. Kyselyn kohdejoukkona oli 4243 suomalaista yritystä, joiden toimintaan liittyi jotenkin digitalisaatio ETLAn aiemmin toteuttaman kyselyn perusteella. Lisäksi kysely lähetettiin kaikille suomalaisille vähintään 50 henkeä työllistäville yrityksille, eli kaiken kaikkiaan 5818 yritykselle. Kyselyyn vastasi 1189 yritystä. Ottaen huomioon, että pk-yritysten osalta kysely rajattiin digitalisaatiota toiminnassaan hyödyntäviin yrityksiin, eivät sen tulokset edusta koko suomen yrityspopulaatiota täysin tarkasti. Kyselytutkimus tarjoaa silti tämän tutkielman kannalta relevantteja tilastoja. On kuitenkin syytä huomioida, että nykyiset luvut massadatan käytöstä suomessa olisivat oletettavasti korkeammat, kuin kyseisessä marraskuussa 2015 toteutetussa kyselyssä. Tämä voidaan melko turvallisesti olettaa, koska tutkimuksessa todettiin selvästi nouseva trendi

massadatan hyödyntämiselle, ja esimerkiksi vuonna 2014 International Data Corporation arvioi maailman massadatateknologioiden ja -palveluiden markkinoille keskimäärin 23 prosentin vuosikasvua (IDC, 2015). Myös IDG Enterprisen kansainvälisessä massadata ja data-analytiikka-kyselytutkimuksessa, joka toteutettiin myös vuonna 2015, havaittiin edellisvuodesta 125 prosentin nousu yrityksissä, joilla on käynnissä datalähtöisiä projekteja (IDG, 2015). ETLAn kyselytutkimuksen ovat toteuttaneet Janne Antikainen, Jarmo Eskelinen, Heli Koski, Tommi Niemi, Mika Pajarinen, Sinikukka Pyykkönen ja Marc de Vries, ja se julkaistiin osana Valtioneuvoston selvitys- ja tutkimustoiminnanjulkaisusarjaa vuonna 2016. Vertaamme kyselytutkimusta paikoittain muihin kansainvälisiin kyselyihin ja tutkimuksiin kattavan kuvan varmistamiseksi.

Massadataa hyödynnetään nykyään liki joka alalla, mutta sen käyttötarkoitukset ja -tavat vaihtelevat alasta ja yrityksen koosta riippuen. Etlan kyselyyn vastanneista suomalaisista yrityksistä 25 prosenttia koki itsensä massadatan hyödyntäjinä. Yleisimmin sitä sovelletaan keskisuurissa ja suurissa, eli yli 50 henkeä työllistävissä yrityksissä. 47 prosenttia, eli liki puolet kyselyyn vastanneista keskisuurista ja suurista suomalaisista yrityksistä hyödynsivät vuonna 2015 massadataa jotenkin toiminnassaan, kun taas vastanneista alle 10 henkeä työllistävistä yrityksistä massadataa hyödynsi vain 21 prosenttia. On huomioimisen arvoista, että kyselyn aihe oli otsikoitu ”Datan käyttö liiketoiminnassa”, ja tämä saattaa olla rajannut vastaajajoukkoa siten, että tästäkin syystä massadatan käyttö näyttäytyy tilastoissa jonkin verran vuoden 2015 Suomen koko yrityspopulaation todellisuutta korkeampana. Vertailuarvona samalle vuodelle e-skills UK-raportissa arvioitiin, että Isossa-Britanniassa suurista yrityksistä massadatan käyttäjiä olisi reilu neljäsosa (e-skills UK, 2014).

Alojen osalta massadatan hyödyntämisen eturintamalla arvioidaan kansainvälisesti olevan rahoitus- ja vakuutus-, sekä energia-alan yritysten. Näillä yrityksillä arvioidaan olevan parhaat mahdollisuudet ja edellytykset luoda arvoa massadatasta (Wolkowitz ja Parker, 2015). Myös ETLAn kyselyssä nämä alat nousivat prosentuaalisesti massadatan hyödyntäjinä kärkipaikoille. Vastanneista neljästätoista rahoitus- ja vakuutusalan yrityksestä kuusi hyödynsi massadataa toiminnassaan, kun taas viidestätoista vastanneesta energia-alan (sähkö-, kaasu- ja lämpöhuoltoa harjoittavasta) yrityksestä massadatan hyödyntäjiä oli kymmenen. Näillä aloilla massadatalle nähtiin kyselyssä myös suurta potentiaalia tulevaisuudessa. Massadatan hyödyntäminen siis näyttäytyi suomessakin näillä aloilla, mutta koska kyselyyn vastanneiden otanta jäi näiden yritysten osalta huomattavan pieneksi verrattuna muihin sektoreihin, ei niistä kannata lähteä vetämään suurempia johtopäätöksiä kyseisille aloille suomessa. Näiden seuraajaksi tilastoissa nousivat logistiikka-ala, jossa nähdään myös paljon potentiaalia massadatalle, sekä informaatio- ja viestintäsektori. Molemmissa massadatan hyödyntäjiä oli hieman yli kolmasosa vastanneista, 34 prosenttia.

Kyselyn perusteella yleisimmin massadatana hyödynnetty data oli peräisin yrityksen myyntijärjestelmistä (61 prosenttia hyödyntäjistä), tai yrityksen www-sivuilta (57-prosenttia hyödyntäjistä). Tämän lisäksi kyselyssä massadatan lähteinä esille nousivat muun muassa avoin data, data kilpailevista tuotteis-

ta, data raaka-aineista ja välituotteista, tuotteiden käyttödata, sekä sosiaalisen median data. Omien lähteidensä hyödyntämisen lisäksi yritykset myös ostavat massadataa toimintansa tueksi. Kyselyyn vastanneista massadatan hyödyntäjistä 35 prosenttia vastasi ostavansa sitä muilta yrityksiltä, ja 12 prosenttia vastasi ostavansa sitä paljon tai hyvin paljon. 15 prosenttia vastasi myyvänsä massadataa toisille yrityksille.

Yleisimmät toiminnot, joihin vastanneet yritykset hyödynsivät massadataa, olivat päätöksenteko (62 prosenttia hyödyntäjistä), myynti ja markkinointi (59 prosenttia hyödyntäjistä), tavaroiden ja/tai palveluiden tuottaminen (56 prosenttia hyödyntäjistä), sekä asiakas- ja markkina-analyysi (55 prosenttia hyödyntäjistä). Massadataa hyödynnettiin eniten siis päätöksenteossa. Se todettiin tämän lisäksi myös toiminnoksi, jolle massadatalla oli suurin merkitys, sillä noin 40 prosenttia massadatan hyödyntäjistä vastasivat hyödyntävänsä sitä paljon tai hyvin paljon päätöksenteossa. Tätä seurasivat toisena ja kolmantena merkityksellisyydessä (hyödynnetään paljon tai hyvin paljon) asiakas- ja markkina-analyysi 39 prosentilla, sekä tuotekehitys 37 prosentilla.

Organisaatioissa on kuitenkin rajoitteita, jotka estävät massadatan hyödyntämistä, ja ETLAn kyselylomake sisälsi kysymyksen myös tämän kartoittamiseen. Tämä osio on hyvin relevantti myös massadatan tietoturvallisuutta arvioitaessa. Kysymys kohdistettiin niillekin yrityksille, jotka eivät hyödynnä toiminnassaan massadataa ollenkaan. Hieman alle puolet näistä yrityksistä eivät nähneet tarvetta massadatan hyödyntämiselle. 52 prosentissa ei-käyttäjien tapauksista tarve kuitenkin tunnistettiin, mutta rajoittavat tekijät löytyivät muualta. Massadatan käsittelyyn ja analysointiin liittyvä tietotaidon puute nähtiin rajoittavana tai estävänä tekijänä noin 30 prosentissa kaikista vastaajayrityksistä. Myös massadatan hyödyntäjien keskuudessa 22 prosenttia koki tietotaidon puutteen rajoittavana tekijänä. Tämä on mielenkiintoinen seikka myös massadatan tietoturvallisuutta harkittaessa, sillä tietoturvallinen datan hallinnointi ja käsittely vaatii myös oman tietotaitonsa (Jonker, W. & Petković, M., 2012). Muita esille nousevia rajoittavia tai estäviä tekijöitä olivat esimerkiksi koetut korkeat kustannukset suhteessa mahdollisiin voittoihin, epäyhteensopivat tietojärjestelmät, jotka eivät esimerkiksi sovellu massadatan keräämiseen, sekä saatavilla olevan datan heikko laatu.

Tietoturvallisuuteen liittyvät syyt olivat kyselytutkimuksen perusteella vähiten koettujen rajoittavien tai estävien tekijöiden joukossa, juridistien syiden ja viranomaisten sääntelyn kanssa. Kaikkien vastanneiden yritysten joukosta vain 17 prosenttia piti tietoturvaa rajoittavana tai estävänä tekijänä massadatan hyödyntämiselle. Tämä kertonee lähinnä siitä, että yritysten luotto olemassa olevien tietoturvaratkaisujen toimivuuteen datavarantojen turvaamiseksi oli ainakin kyselytutkimuksen aikaan noussut suhteellisen korkealle tasolle, sillä massadata kuitenkin todistetusti esittää tietoturvariskinsä (N. Chaudhari ja S. Srivastava, 2016). Tätä selitystä vahvistavat esimerkiksi IDG Enterprisen kansainvälisten data-aiheisten yrityskyselyiden tulokset. Vuonna 2014 IDG:n kansainväliseen verkkokyselyyn vastanneista yritysten edustajista 44 prosenttia piti yrityksen silloisia tietoturvaratkaisuja riittävinä yrityksen datavarantojen tur-

vaamiseksi, kun tämä luku oli noussut vuonna 2015 66 prosenttiin (IDG, 2015). Tämän perusteella kyselytutkimuksen aikana oli siis havaittavissa selvää nousua yritysten luotossa tietoturvaratkaisuihin myös globaalilla tasolla. Organisaatioiden on silti tärkeää olla tietoisia mahdollisista tietoturvaan liittyvistä riskeistä, koska näiden tiedostamattomuus nostaa riskin toteutumisen tasoa.

Tietosuoja nähtiin ETLAn kyselytutkimuksessa massadatan hyödyntämistä rajoittavana tai estävänä tekijänä 21 prosentissa kaikista vastanneista yrityksistä. Massadatan esittäessä hyvin selkeitä haasteita tietosuojan osalta (Zhang, 2018), vaikuttaa tämä luku suhteellisen pieneltä. Tätä selittänee osaksi se, että kyselytutkimus on toteutettu useamman vuoden ennen GDPR:n astumista voimaan, ja ennen kuin EU antoi itse asetuksen huhtikuussa 2016. Näin ollen suuri osa nykyisestä tietosuojakeskustelusta ei ollut noussut vielä pinnalle.

2.2 Massadatan hyödyntämisen käytännöt ja menetelmät organisaatioissa

Tässä luvussa selvennetään tarkemmin, miten organisaatiot käytännössä keräävät, käsittelevät ja hyödyntävät massadataa toiminnassaan. Massadataa hyödynnetään sekä yksityisissä että julkisissa organisaatioissa erilaisiin tarkoituksiin (Antikainen et al., 2016). Käsittelemme seuraavaksi näiden organisaatioiden soveltamia menetelmiä datan keräämiselle, säilömiselle ja käsittelylle, antaaksemme lukijalle paremman kuvan niihin liittyvien tietoturvariskien luonteesta ja merkityksestä.

Data tulee nykyään enenevissä määrin lähteistä, joiden myötä sen rakenne on luonteeltaan sekavaa ja epäjärjestynyttä. Yksi tyypillinen tällainen lähde ovat esimerkiksi erilaiset koneet ja sensorit, jotka ovat kasvava massadatan lähde esineiden internetin, eli IoT-ratkaisujen yleistyessä (Cai, Xu & Jiang, 2017). IoT:llä viitataan kattavaan verkkoon älykkäitä objekteja/laitteita, joilla on kapasiteetti jakaa informaatiota, dataa ja resursseja, ja jota reagoivat ja tekevät toimenpiteitä ympäristössä tapahtuviin muutoksiin ja tapahtumiin (Makadam, Ramaswamy & Tripathi, 2015). Tämän lisäksi massadatan lähteinä esiintyy monia muita suuria avoimen ja yksityisen datan lähteitä, kuten verkkosivuja ja järjestelmiä, joita edellisessä osiossakin eriteltiin.

Yli 4 miljardia ihmistä maailmanlaajuisesti tuottaa dataa mobiililaitteillaan ja tietokoneillaan selatessaan internetiä ja käyttäessään erilaisia sovelluksia (Russom, 2011). Tästä syntyvä data on massiivista ja hyvin epäjärjestäytynyttä. Internettiä ja sitä selaavia ihmisiä voidaankin pitää yhtenä keskeisimmistä massadatan tuottajista. Tätä dataa on kuvailtu myös ”ihmisten tuottamaksi dataksi”, ja se koostuu esimerkiksi kuvista, videoista, rahallisista transaktioista verkkokaupoissa, liikkumistiedoista, sosiaalisen median, kuten Facebookin, päivityksistä, ja muusta sosiaalisen median datasta (Douglas, 2012). Organisaatiot keräävät tätä dataa, ja jalostavat siitä erilaisilla menetelmillä hyödyllistä tietoa itselleen. Tämä hyöty voi olla esimerkiksi parempaa ymmärrystä jostakin popu-

laatiosta, ja sillä voidaan tavoitella esimerkiksi etua markkinoinnin kohdentamisessa. Massadatan hyödyntäminen markkinointiin on yksi sen hyvin yleinen sovelluskohde (Erevelles, Sunil, Fukawa, & Swayne, 2016).

Vanhat työkalut eivät riittäneet tällaisten datamäärien keräämiseen ja varastoimiseen, eikä niillä pystytty käsittelemään massadataa mielekkäässä ajassa, tai jalostamaan siitä relevanttia informaatiota. Uusien datateknologioiden ja -työkalujen myötä tästä on kuitenkin tullut mahdollista. Massadatan käsittelyyn ja keräämiseen suunnitellut työkalut on suunniteltu tekemään datan hallinnoimisesta ymmärrettävää, luotettavaa, turvallista ja hallittavaa. (Khan, Yaqoob, Abaker, Hashem, Inayat, Kamaleldin, Ali, Alam, Shiraz, Gani, 2014). Tämän teknologisen kehityksen myötä suuria datamääriä pystytään nykyään käsittelemään ilman ”supertietokoneita” tai liian korkeita kustannuksia.

Joitakin yleisesti hyödynnettyjä teknologioita, joita voidaan käyttää massadatan hallinnoimiseen, ovat esimerkiksi Google BigTable, Simple DB, sekä monet muut Not Only SQL (NoSQL) -tietokantaratkaisut (M. Chen, S. Mao, and Y. Liu, 2014). NoSQL on käsite, jolla kuvataan perinteisestä relaatiomallista poikkeavia tietokantaratkaisuja. Nämä NoSQL-tietokantaratkaisut ovat sovellettavissa massadatan käsittelyyn, kun taas perinteiset relaatiotietokantaratkaisut ovat tyypillisesti riittämättömiä. Tämä johtuu siitä, että massadata on yleensä liian runsasta, rakenteetonta, ja hajanaista ollakseen sijoitettavissa relationaalisiin tauluihin, ja NoSQL-tietokantoja on helpompi skaalata vastaamaan muuttuvia vaatimuksia (Gudivada et al., 2014). Relaatiotietokannoista on kuitenkin kehitetty massadatan käsittelyä varten uudistettu malli, jota kutsutaan NewSQL:ksi. NewSQL-tietokannat pyrkivät pärjäämään suorituskyvyssä NoSQL-tietokannoille, mutta tarjoamaan silti kyselyitä SQL-kielellä. (Strohbach et al., 2016.)

Tämän lisäksi yritykset tarvitsevat lisäratkaisuja suurten datamäärien varastoimiseen, käsittelyyn ja analysointiin reaaliajassa, koska massadata eroaa muista datajoukoista siten, että sitä ei voida varastoida yhdelle laitteelle, kuten tavanomaiselle tietokoneen kovalevylle. Tästä syystä organisaatiot joutuvat hyvin usein varastoimaan hallinnoimansa massadatan useammalle palvelimelle, ja tarvitsevat erityisiä työkaluja datan käsittelemiseen yli palvelimien. Joitakin tähän yleisesti käytettyjä työkaluja ovat esimerkiksi Hadoop ja MapReduce (Khan et al., 2014).

Hadoop on Apache Software Foundationin kehittämä avoimen lähdekoodin ohjelmisto suurien datamäärien käsittelyyn, ja projektin voidaankin sanoa olevan kehitetty massadatan prosessointiin. Joitakin massadatan sovelluksia, joihin Hadoopia on käytetty, ovat esimerkiksi roskapostin tunnistaminen, hakukonealgoritmit, ja sosiaalisen median alustoiden suosittelualgoritmit. Esimerkiksi Yahoo soveltaa massadataa Hadoopin avulla palveluidensa toiminnallisuuksiin, kuten edellä mainittuihin hakukonealgoritmeihin ja roskapostisuodattimiin. Kesäkuussa 2012 laskettuna, Yahoo oli soveltanut Hadoopia 42 000 palvelimeen neljässä eri datakeskuksessa. (M. Chen, S. Mao, and Y. Liu, 2014.) MapReduce puolestaan on ohjelmointityökalu, jonka avulla Hadoopia voidaan soveltaa. Sen avulla voidaan käsitellä dataa hallittavasti ”klustereina”,

joihin kuuluu dataa useilta eri palvelimilta. Tällaiset työkalut ovat tehneet massadatan hyödyntämisestä mahdollista, sillä niiden avulla pystytään onnistuneesti käsittelemään suuria datamääriä hallittavasti, kustannustehokkaasti, ja mielekkäässä ajassa (A. O'Driscoll, J. Daugelaite, and R. D. Sleator, 2013).

Massadatan elinkaari organisaatiossa voidaan katsoa koostuvan seuraavista vaiheista (Khan et al., 2014):

1. datan kerääminen, suodattaminen ja lajittelu
2. datan analysointi
3. datan varastointi, jakaminen ja hyödyntäminen
4. datan uudelleenkäyttäminen

Elinkaaren aikana on pyrkimyksenä jalostaa niin kutsutusta raakadatasta organisaation toimintaan sovellettavaa hyödyllistä informaatiota. Prosessi lähtee liikkeelle datan keräämisestä. Tässä elinkaaren mallissa käsitellään siis organisaatioita, jotka keräävät itse massadataa. On huomioimisen arvoista, että jotkin organisaatiot vain ostavat ja hyödyntävät toisten keräämää massadataa, jota on usein jo jalostettu, jolloin tämä elinkaaren malli ei ole tilanteeseen täysin sovellettavissa, ja prosessi on jokseenkin yksinkertaisempi. Kerättävä data voi koostua esimerkiksi edellä mainitusta ”ihmisten tuottamasta datasta”, kuten hakukoneisiin kirjoitetuista hakulauseista, verkkokauppojen transaktioista, ja sosiaalisen median päivityksistä (Douglas, 2012). Raakadatan keräämiseen eri lähteistä on olemassa useampia tekniikoita, joita organisaatiot soveltavat. Näitä tekniikoita ovat esimerkiksi (Khan et al., 2014):

- Lokitiedostot, joihin palvelimet voivat kerätä tietoa esimerkiksi klikkauksista ja muista käyttäjätiedoista.
- Erilaiset sensorit, jotka voivat mitata fyysisiä ominaisuuksia.
- ”Web crawlerit”, eli hakurobotit ja muut internetiä selaavat työkalut, jotka keräävät tietoa automatisoidusti suuresta määrästä nettisivuja.
- Evästeet, joita ihmisten selaimet keräävät eri palvelimilta, ja joiden avulla käyttäjän vierailemat nettisivut taas saavat dataa itselleen.

Tämä data täytyy suodattaa ennen tallentamista. Havainnollistamisen vuoksi voidaan ajatella esimerkiksi valvontakameroista kerättävää dataa. Massiivinen määrä videomateriaalia ei ole hyödyllistä, silloin kun kuvassa ei tapahdu mitään. Massadatan monipuolisen luonteen vuoksi suodattaminen on haasteellinen prosessi. Kerättyä dataa tulee myös usein yhdistää toiseen dataa, jotta siitä saadaan suurin mahdollinen arvo irti, ja se täytyy tallentaa järjestäytyneessä formaatissa. Järjestäytynyt formaatti on tärkeä ominaisuus kyselyiden mahdollistamiseksi. (Toshniwal, Raghav, Kanishka Ghosh Dastidar & Asoke Nath, 2015.) Modernit datateknologiat luokittelevat ja suodattavat tätä dataa niille annetuilla parametreillä ennen tallentamista.

Seuraavaksi kerätty data täytyy analysoida, jotta siitä voidaan jalostaa toiminnan kannalta hyödyllistä informaatiota. Massadatan tapauksessa analy-

sointi tapahtuu usein reaaliajassa rinnakkain keräämisen kanssa, koska kaikkea massadataa ei haluta säilöä organisaation datavarantoihin pysyvästi. Prosessi on useassa tapauksessa haastava massadatan kompleksisuuden ja suuren määrän vuoksi. Se vaatii esimerkiksi paljon laskentatehoa sitä suorittavilta palvelimilta, ja siihen liittyy useita muitakin haasteita. (A. Labrinidis and H. Jagadish, 2012). Datan analysoinnilla voidaan katsoa olevan kaksi ensisijaista päämäärää; ymmärtää johdonmukaisuudet ja syy-seuraussuhteet datan ominaisuuksista, ja kehittää tehokkaat ja tarkoituksenmukaiset tiedonlouhinnan menetelmät relevantin informaation jalostamiseksi (Fan & Liu, 2013). Massadatan analysointi vaatii erikoistuneet prosessinsa ja työkalunsa muun muassa sen nopeasti lisääntyvän luonteen ja suuren koon vuoksi. Tämän lisäksi haasteita esittää se, että massadata sisältää tyypillisesti dataa monessa eri formaatissa (Michael & Miller, 2013). Analysointiin sovelletaan menetelminä muun muassa tiedonlouhintaa, visualisointia, tilastotieteellistä analyysiä, ja koneoppimista. Esimerkiksi tiedonlouhinnalla, tunnettu yleisemmin termillä ”Data mining”, voidaan automatisoidusti löytää hyödyllisiä johdonmukaisuuksia suurista datajoukoista. Menetelmään yhdistetään usein koneoppimisen käytänteitä, ja sillä pystytään jalostaa epäorganisoituneesta ja sekavasta datajoukosta hyödyllistä informaatiota. Tiedonlouhinnan saralta löytyy useita massadatan analysointiin soveltuvia työkaluja. (Wu, Xindong, et al, 2013)

Analysoinnista saatu jalostettu ja lajiteltu, toiminnan kannalta hyödyllinen informaatio säilötään oleellisilta osin organisaatiossa tulevaa käyttöä varten, ja jaetaan se sitä tarvitseville henkilöille hyödynnettäväksi. Datan varastoimisessa pyritään luotettavuuteen, turvallisuuteen, sekä hyvään saatavuuteen ja käytettävyyteen (Khan et al., 2014). Massadatan varastointimenetelmät ovat myös sen tietoturvallisuuden kannalta keskeinen tekijä. Tässä on oleellista huomata ero niiden organisaatioiden, jotka keräävät itse massadataa, ja niiden organisaatioiden, jotka vain hyödyntävät sitä välillä. Massadataa keräävät organisaatiot joutuvat kiinnittämään huomattavasti enemmän huomiota datan varastointimenetelmiin, kun taas pelkästään sitä ostamalla hyödyntävät organisaatiot pääsevät tässä helpommalla, sillä itse säilöttävän datan määrä on tässä tapauksessa pienempi (Liang, Fan, et al, 2018).

Yhden mallin mukaan massadataa tukevat varastointiratkaisut voidaan jakaa kahteen pääkategoriaan; erillisiin varastointiratkaisuihin (Storage system for large data), ja jaettuihin varastointiratkaisuihin (Distributed storage system). Erilliset varastointiratkaisut voidaan edelleen jakaa DAS-ratkaisuihin (Direct attached storage, eli suoraan kiinnitetty varastointi), ja NS-ratkaisuihin (Network storage, eli verkkotallennus). DAS-ratkaisuissa yhteen tietokoneeseen tai palvelimeen on kiinnitetty useampi fyysinen kovalevy datan varastoimiseksi. Tämä soveltuu rajoitetusti massadatan hallinnoimiseen, sillä datan lisääntyessä nopeasti, on ratkaisun skaalaaminen haasteellista. NS-ratkaisuissa palvelin on liitetty suoraan organisaation tietoverkkoon, ja asianomaiset pääsevät siihen käsiksi esimerkiksi VPN-yhteyden avulla. Jaetuissa varastointiratkaisuissa puolestaan useita palvelimia liitetään toisiinsa jaetussa tietoverkossa, ja niitä yhdistää yhteinen datan varastointi- ja hallinnointijärjestelmä, esimerkiksi pilvipalve-

luympäristö. Nämä palvelimet sijaitsevat yleensä eri paikoissa. Tämä on suurimpien massadatan käsittelijöiden yleisesti suosima ratkaisu. Massadatan hallinnoiminen tapahtuu nykyään tyypillisesti siis pilvipalveluympäristössä. (Khan et al., 2014) Datavolyymien kasvaessa suuremmiksi ja suuremmiksi, nousee niiden varastointi keskeiseksi keskustelunaiheeksi. Seuraavassa luvussa perehdytään tarkemmin massadatan yleisiin varastointimenetelmiin, ja käsitellään tarkemmin siihen liittyviä ongelmia.

Lopulta, jotta kerätystä ja jalostetusta massadatasta saadaan suurin mahdollinen arvo irti, on sen uudelleen käyttäminen suositeltavaa, mikäli mahdollista. Tutustumalla uudelleen jo kerättyyn dataan, on mahdollista löytää siitä uutta arvoa, ja soveltaa sitä uusiin käyttötarkoituksiin. Massadata on luonteeltaan sellaista, että kaikkia sen mahdollisia sovelluksia ei yleensä ole välittömästi havaittavissa, ja tästä syystä dataan on suositeltavaa tutustua uudelleen, sillä siitä saatetaan pystyä jalostamaan aiemmin huomaamatonta arvoa. (Hurwitz et al., 2013.)

Massadatan käyttötarkoituksia ja tavoitteita organisaatioille on tuotu nyt kattavasti esille tässä ja edeltävässä osiossa. Sovellusmahdollisuuksia onkin liki rajattomasti, ja tutkielman aiheen rajaamisen vuoksi niihin ei voida paneutua rajattomasti. Massadata ei ole kuitenkaan pelkästään yksityisten organisaatioiden työkalu, ja sitä sovelletaan moniin tarkoituksiin myös julkisissa organisaatioissa. Edellä esitetyt massadatan soveltamisen käytännöt ja pelisäännöt ovat pitkälti niiden tapauksessa hyvin samanlaisia, mutta tyypilliset käyttötarkoitukset ja -tavoitteet näyttäytyvät välillä jokseenkin erilaisina. Julkisissa organisaatioissa ja palveluissa massadatalle kehitetään sovelluksia ja ratkaisuja esimerkiksi terveydenhuoltoon sairauksien ennustamiseen ja hoidon suunnittelemiseen. Tämän lisäksi keskeinen painopiste on myös julkisessa liikenteessä älykkäiden ratkaisujen saralla, esimerkiksi liikenteen aikatauluja ja reittejä suunniteltaessa (Antikainen et al., 2016).

3 Massadatan tietoturvariskit ja niiden hallitseminen organisaatioissa

Tässä luvussa käsitellään edellä selvennetyn massadatan hyödyntämisen organisaatioille aiheuttamia tietoturvariskejä ja -ongelmia. Luvun on tarkoitus vastata tutkielman toiseen tutkimuskysymykseen ”Millaisia tietoturvaongelmia massadatan hyödyntäminen aiheuttaa?”. Tietoturvallisuudella tarkoitetaan järjestelyjä, joilla pyritään varmistamaan tiedon (Lundgren, Björn, & Möller, 2019):

- käytettävyys, eli tiedot ovat käyttäjän saatavilla tarpeen mukaisesti
- eheys, eli tiedon sisältöä tai rakennetta ei ole muutettu tahallisesti tai tahattomasti
- luottamuksellisuus, eli tietoa voivat käsitellä vain henkilöt, joilla on siihen oikeus
- kiistämättömyys, eli käyttäjän tekemää tekoa järjestelmässä ei voida luotettavasti kiistää
- todennettavuus, eli käyttäjä voidaan tunnistaa luotettavasti luonnolliseksi ja/tai oikeushenkilöksi
- tunnistettavuus, eli käyttäjä voidaan tunnistaa käyttäjätunnuksen perusteella

Massadatan keräämisestä, säilömisestä, ja käsittelystä aiheutuu organisaatioille monia tietoturvariskejä, jotka voivat toteutuessaan johtaa ongelmiin, jos niihin ei olla varauduttu (CSA, 2013). Digitalisoitumisen mukanaan tuomat työkalut, kuten massadata ja sen soveltaminen halutaan usein nähdä vain taloudellisen kasvun ja tehokkuuden mahdollisuuksien kautta, eikä niiden todellista kokonaisvaltaisuutta aina nähdä riittävästi. Tästä syystä organisaatioiden on tärkeää olla perillä myös näistä riskitekijöistä. Luku jaetaan kolmeen osaan. Tietosuojan ja GDPR:n ollessa niin keskeinen osa massadatan tietoturvaongelmia, omistetaan yksi osa niille. Toisessa osassa keskitytään massadatan tietoturvaongelmiin yleisemmällä tasolla, ja kolmannessa käsitellään tiiviisti, miten organisaatioissa ollaan tietoisia massadatan aiheuttamista tietoturvahaasteista, ja kuinka niihin ollaan varauduttu.

3.1 Tietosuoja ja GDPR massadatan kontekstissa

Ihmisistä ja heidän toiminnastaan kerätään nykyään dataa enemmän kuin koskaan aiemmin. Tämä esittää sekä hyötyjä, että haittapuolia. Toisaalta data tuo tehokkuutta organisaatioiden toimintaan ja arjen sujuvuuteen, mutta toisaalta ihmisten tuottaman datan kerääminen ja käsittely laajalla mittakaavalla saattaa johtaa myös tahattomaan, pahimmassa tapauksessa arkaluontoistenkin henkilötietojen keräämiseen, tai näiden tietojen vuotamiseen. Massadatan luonteen vuoksi tätä on ajoittain hyvin vaikea hallita, ja henkilötiedot saattavat vuotaa vääriin käsiin, vaikka organisaatio olisikin saanut rekisteröidyltä luvan niiden hallinnoimiseen. (Zhang, 2018). Tässä alaluvussa käsitellään massadatan tietosuojahaasteita. Käydään ensiksi läpi aiheen kannalta relevantti GDPR-tietosuoja-asetus, ja sen implikaatiot aiheelle.

3.1.1 GDPR-tietosuoja-asetus

Yksityisyydensuojan hallinta on tällä hetkellä yksi keskeisistä massadataa koskettavista haasteista, ja se on noussut entistä keskeisempään asemaan vuonna 2018 voimaan astuneen GDPR-tietosuojasäädöksen myötä. Käymme tässä kappaleessa GDPR:n konseptin ja juridisen puolen tiiviistettynä läpi, sillä se on keskeinen osa massadatan tietoturvahaasteiden ymmärtämistä. Kappaleen ensisijaisena lähteenä käytetään suoraan itse GDPR-tietosuoja-asetusta viimeisimmässä vuonna 2018 päivitetyssä versiossaan (GDPR, 2018). Jos esitetään väittämä, joka ei ole poimittu itse tietosuoja-asetuksesta, viitataan siihen erillisellä lähteellä. GDPR tulee sanoista General Data Protection Regulation. Se on henkilötietojen käsittelyä sääntelevä laki, jonka Euroopan Unioni asetti vuonna 2016, ja se astui voimaan kaikissa EU-maissa 25. toukokuuta 2018. Henkilötiedoilla siinä tarkoitetaan kaikkia tunnistettuun tai tunnistettavissa olevaan henkilöön liitettäviä tietoja. Tunnistettavissa olevana henkilönä pidetään luonnollista henkilöä, joka voidaan suoraan tai epäsuoraan tunnistaa erilaisten kerättävien tunnistetietojen, kuten ”nimen, henkilötunnuksen, sijaintitiedon, verkkotunnistetietojen, taikka yhden tai useamman hänelle tunnusomaisen fyysisen, fysiologisen, geneettisen, psyykkisen, taloudellisen, kulttuurillisen tai sosiaalisen tekijän perusteella”. Nämä ovat määritelmällisesti henkilötietoja, ja tässä tutkielmassa käytetään samaa määritelmää henkilötiedoista puhuttaessa. Tietosuojalla puolestaan tarkoitetaan näiden henkilötietojen käsittelyä siten, että ne pysyvät henkilön itsensä hallinnassa, eivätkä ajaudu henkilön tahtomatta sellaisen tahon käsiin, kenellä ei ole lainvoimaista perustetta tietojen hallinnoimiseen.

GDPR:n ensisijainen tarkoitus on antaa ihmisten henkilötiedoille aikaisempaa parempi suoja, ja lisätä tietosuojaoikeuksia, kuten keinoja hallita ja tarkastella, mitä dataa itsestä kerätään. Tämän lisäksi tavoitteena on ollut vastata uusiin digitalisaatioon ja globalisaatioon liittyviin tietosuojakysymyksiin, ja yhtenäistää tietosuojasääntelyn periaatteita kaikissa EU-maissa. GDPR:n ensimmäinen artikla, joka voidaan nähdä myös sen ”punaisena lankana”, sanoo, et-

tä ”Luonnollisten henkilöiden suojelu henkilötietojen käsittelyn yhteydessä on perusoikeus.” Asetuksen jokaista pykälää ei pystytä tutkielman mittasuhteissa käymään läpi, joten pyritään tiivistämään GDPR:n keskeinen tarkoitus ja periaatteet mahdollisimman tehokkaasti. GDPR:n tietosuojaperiaatteiden mukaan henkilötietoja on:

- käsiteltävä lainmukaisesti, asianmukaisesti, ja läpinäkyvästi
- käsiteltävä luottamuksellisesti ja turvallisesti
- kerättävä ja käsiteltävä tiettyä, nimenomaista ja laillista tarkoitusta varten
- kerättävä vain tarpeellinen määrä käsittelyn tarkoitusta varten
- päivitettävä aina tarvittaessa. Epätarkat ja virheelliset henkilötiedot tulee korjata tai poistaa mahdollisimman pian.
- säilytettävä sellaisessa muodossa, josta rekisteröity on tunnistettavissa ainoastaan niin kauan, kuin on tarpeen tietojenkäsittelyn tarkoituksen saavuttamiseksi.

GDPR:n mukaan organisaatio siis tarvitsee lainvoimaisen perusteen henkilötietojen käsittelylle. Tämä peruste voi olla esimerkiksi rekisteröidyn suostumus, erillinen sopimus, tai rekisterinpitäjän lakisääteinen velvoite tietojen käsittelyyn, joka voi perustua esimerkiksi elintärkeiden etujen suojaamiseen. Asetuksen myötä ihmisille on myös tullut useita uusia oikeuksia tietosuojaan liittyen. Rekisteröidyllä on oikeus esimerkiksi tietää, mitä henkilötietoja organisaatio hänestä hallinnoi, mihin tarkoitukseen näitä tietoja käsitellään, pyytää virheellisten ja puutteellisten henkilötietojen korjaamista, vastustaa henkilötietojen käsittelyä, tai pyytää kokonaan niiden poistamista. Yksi massadatan ominaisuuksista on monipuolisuus. Datan diversiteetti esittää merkittäviä haasteita yhtä monipuolisten turvallisuusvaatimuksien vuoksi, joista GDPR on hyvä esimerkki (Strohbach et al., 2016). Organisaatiot joutuvat ilmoittamaan useita tietoja henkilötietojen keräämisen yhteydessä:

- organisaation nimen ja yhteystiedot
- mitä henkilötietoja organisaatio aikoo kerätä
- mihin tarkoitukseen henkilötietoja aiotaan käsitellä
- millainen oikeusperuste organisaatiolla on tietojen käsittelyyn
- kuinka kauan tietoja säilytetään
- keille kaikille henkilötiedot jaetaan
- mitä oikeuksia henkilöllä on, kuten peruuttaa suostumukseen perustuva valtuutus henkilötietojen käsittelyyn missä vaiheessa tahansa.

GDPR:n periaatteisiin liittyy myös ”tietosuojalähtöinen suunnittelu”. Sen mukaan organisaatioiden tulisi ottaa yksityisyydensuoja huomioon suunnitellessaan, implementoidessaan ja operoidessaan mitään teknologiaa, joka käsittelee henkilötietoja. Ennen GDPR:ää, vastuu jäi alustan käyttäjälle huolehtia, mitä dataa hän luovuttaa organisaatiolle. Käyttäjän tuli esimerkiksi omatoimisesti

muuttaa alustan perusasetuksia, tai kytkeä sijaintitietojen jakaminen pois, jotteivat hänen henkilötietonsa ajaudu organisaation käsiteltäväksi. GDPR:n tietosuojalähtöisen suunnittelun periaate vaatii organisaatioita implementoimaan tietosuojastandardeja alustoihinsa, ja tarjoamaan niitä käyttäjille oletuksena. Näin ollen GDPR-asetus on siirtänyt huomattavan määrän vastuusta käyttäjältä organisaatiolle tietosuojan ylläpitämisessä.

3.1.2 Tietosuojan toteutumisen haasteet massadatan kontekstissa

Massadatan ja internetin aikakautena organisaatioiden hyödyntämät datajoukot ovat kasvaneet suuremmiksi ja monimutaisemmiksi kuin koskaan aiemmin, ja tämän myötä myös niiden luotettava ja hallinnoitava käsittely on hankaloitunut. Massadatan ollessa suurilta osin ihmisten luomaa dataa, päätyy sekaan väistämättä henkilötietoja, ja tämä aiheuttaa organisaatioille merkittäviä haasteita. Datan keräämisen yhteydessä tulee välttää aiheetonta henkilötietojen keräämistä, eikä luvallakaan kerättyjä henkilötietoja saa vuotaa eteenpäin (Zhang, 2018).

Aiemmassa kappaleessa käsitellyssä ETLAn vuonna 2016 julkaisemassa massadata-kyselytutkimuksessa tunnistettiin massadatan merkittävimmäksi riskiksi henkilötietoriskit. Massadatan sekaan päätyy usein luottamuksellisia henkilötietoja, ja niiden poistamisen tai anonymisoinnin haasteellisuus vaihtelee. Massadata on sekavan ja nopeasti lisääntyvän luonteensa vuoksi sellaista, että organisaatio saattaa tahtomattaankin kerätä tai päätyä käsittelemään henkilötietoja sitä soveltaessa. Ennen kaikkea ongelmaksi muodostuu se, että jopa oikeaoppisesti anonymisoidusta datasta voidaan tunnistaa yksilöitä tiedonlouhinnan menetelmillä. Myös datasta, joka ei välttämättä sisällä näennäisesti ollenkaan henkilötietoja, saattaa louhinnan kautta olla mahdollista muodostaa sellaisia. Datan anonymisointi osoittautuu ajoittain siis riittämättömäksi keinoksi henkilötietojen salaamiseksi, sillä suuri määrä dataa mahdollistaa henkilöiden uudelleen tunnistamisen. (Strohbach, Daubert, Ravkin & Lischka, 2016). Mitä enemmän näennäisesti epäolennaista tietoa on saatavilla eri lähteissä, sitä helpommin ulkopuolinen taho pystyy niin sanotusti ”yhdistämään pisteet” ja saamaan haluamansa tiedon jalostettua (Jonker, W. & Petković, M., 2012). Tutkimuksella on osoitettu, että julkisesti avoimesta informaatiosta on pystytty arvaamaan yksittäisen henkilön sosiaaliturvatunnus data-analytiikan menetelmiä hyödyntämällä (Acquisti & Gross 2009). Datanlouhinnan massadatalle esittämät tietoturvahaasteet havainnollistuvat hyvin henkilötiedoissa, mutta samalla periaatteella on louhittavissa massadatan seasta myös muutakin arkaluontoista dataa. Tämä voi olla esimerkiksi yritysten liiketoiminnan kannalta sensitiivistä tietoa. Datanlouhinnan esittämien haasteiden vuoksi massadatavarantojen osalta on erityisen tärkeää olla huolellinen käyttäjien oikeuksien ja pääsynhallinnan suhteen, mikä voi olla ajoittain haasteellista.

Ihmiset luovuttavat erinäisiin palveluihin huomaamattaan tietosuojan toteutumisen kannalta kriittistä dataa. Usein tämä tapahtuu myös tietoisesti. Kuluttajat ja erilaisten palveluiden asiakkaat luovuttavat esimerkiksi luottokortti-

tietonsa verkkokauppoihin, kertovat apteekissa asioidessaan mitä lääkkeitä tarvitsevat, ja jakavat henkilötunnuksensa erilaisille palveluntarjoajille.

Tämä on merkittävä haaste massadatan kannalta myös siksi, koska monessa maassa kansalaisten nykyluottamus henkilötietojen asianmukaiseen käsittelyyn ja salassapitoon on jo valmiiksi heikolla tasolla (Antikainen et al., 2016). Datan anonymisointi on sen heikkouksista huolimatta yksi parhaista keinoista käsitellä massadatan tietosuojaan liittyviä ongelmia. Oikeaoppisesti anonymisoitua data ei esimerkiksi pidetä edellä esitellyssä GDPR-tietosuojasäädöksessä henkilötietoina ollenkaan, ja sitä saa hyödyntää toiminnassaan vapaasti (GDPR, 2018). Anonymisoinnin sijasta data voidaan myös pseudonymisoida. GDPR:n mukaan pseudonymisoinnilla tarkoitetaan datan käsittelyä siten, että henkilötietojen perusteella ei voida tunnistaa rekisteröityä ilman erillään pidettyä lisätietoa. Menetelmässä vaaditaan, että lisätieto pidetään huolellisesti erillään henkilötiedoista. Pseudonymisointi ei ole henkilötietojen turvaamisen kannalta yhtä tehokasta kuin tietojen anonymisointi, mutta se on parempi ratkaisu joissain tapauksissa, joissa organisaatio joutuu säilyttää henkilötietoja siten, että se pystyy kuitenkin tunnistamaan rekisteröidyn niistä tarvittaessa. Näissä tapauksissa organisaatio voi pseudonymisoimalla huolehtia siitä, etteivät epäasianomaiset voi tunnistaa rekisteröityä datasta. Pseudonymisointiin liittyy myös omat haasteensa. Organisaation prosessit ja tekniset toiminnot tulee olla huolellisesti suunniteltuja sen osalta, jotta avaintieto pysyy oikeasti erillään pseudonymisoidusta datasta, eivätkä väärät henkilöt saa molempia käsiinsä. (Mourby, Mackey, Elliot, Gowans, Wallace, Bell, Smith, Aidinlis & Kaye, 2018)

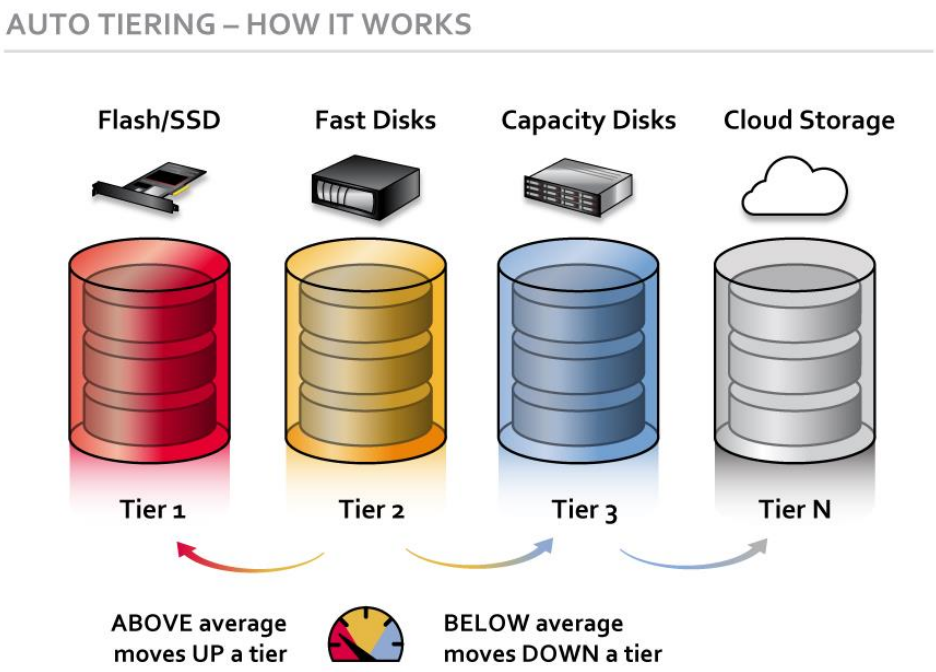
Paitsi että massadatan keräämisen yhteydessä on helppo huomaamatta kerätä aiheettomasti henkilötietoja, nämä tiedot saattavat vuotaa vääriin käsiin osana massadataa myös varastoinnin, siirtämisen, tai käytön aikana, vaikka henkilötietojen kerääminen itsessään olisikin toteutettu laillisesti. Esimerkiksi Facebookin tapauksessa, jota pidetään yhtenä maailman johtavista massadatan hallitsijoista, Skull Security -yrityksen tutkija Ron Bowes onnistui keräämään ihmisten henkilötietoja 2.8 Gigan edestä informaationhankintatyökalulla. Kyseiset käyttäjät olivat jättäneet manuaalisesti piilottamatta omat henkilötietonsa, ja tällainen tapahtuu hyvin helposti, jos palvelua ei ole suunniteltu tietosuojalähtöisesti (Chen, Mao & Liu, 2014). Voidaan myös sanoa, että henkilötiedot houkuttelevat hyökkääjiä niiden arkaluontoisuuden vuoksi. Henkilötietojen esiintyminen yhdessä massadatavarantojen suuren koon kanssa luovatkin hyökkääjille niistä erittäin mielekkään kohteen.

3.2 Massadatan tietoturvaongelmat

Tässä alaluvussa käsitellään massadatan tietoturvaongelmia yleisemmällä tasolla, ja keskitytään niiden kokonaiskuvaan. Edellä esitetyt yksityisyydensuojan aiheuttamat haasteet ovat keskeinen osa massadatan tietoturvaongelmia, mutta se esittää alana myös muita tietoturvariskejä organisaatioille. Käsitellään ensisijaisesti massadatan keräämisestä, varastoinnista ja hallinnoimisesta aiheutuvia tietoturvaongelmia.

Massadatan varastointi on sen monimutkaisen luonteen vuoksi osaltansa haavoittuvaista. Massadatan varastointiin sovelletaan mitä yleisimmin NoSQL-ratkaisuja tavanomaisten relaatiotietokantojen sijasta. NoSQL-tietokantoihin ja tavanomaisiin relaatiotietokantojen hallintajärjestelmiin (RDBMS) kohdistuvat samantyyppiset tietoturvariskit ja -ongelmat (Winder, 2012). NoSQL-tietokannoista kuitenkin puuttuu monet turvallisuustoimenpiteet, jotka ovat relaatiotietokantojen hallintajärjestelmässä oletusarvoisesti kytkettyinä. Näitä toimenpiteitä ovat esimerkiksi sensitiivisen datan salaus, prosessien ajaminen hiekkalaatikossa, syötteiden validointi, ja vahva käyttäjien autentikointi. (Okman et al., 2011) NoSQL-tietokantaratkaisujen turvallisuuden vahvistaminen on keskeinen aihe massadatan aiheuttamien tietoturvaongelmien ratkaisemisessa.

Lisää massadatan varastointiin liittyviä turvallisuushaasteita esittää datan automaattinen jakelu palvelimien ja muiden tallennusmedioiden välillä. Massadatavaranto ei usein mahdu yhdelle palvelimelle, ja datan hallinnoijat luovuttavat kontrollin datan jakelusta ja säilömisestä algoritmeille kustannusten pienentämiseksi. Tästä puhutaan yleensä englanninkielisellä käsitteellä ”tiering” tai ”auto tiering”. Data liikkuu dynaamisesti tallennuskohteesta toiseen, usein sen tärkeyden perusteella. Datan sijaintia, liikkeitä ja muutoksia datassa joudutaan seuraamaan transaktiolokien avulla. Datan automaattinen jakelu tulee suunnitella huolellisesti, jotta voidaan välttää sensitiivisen datan ajautuminen huonosti suojattuihin kohteisiin. (CSA, 2013) Massadatan varastointi klustereina useassa eri kohteessa johtaa myös siihen, että organisaatioiden on haasteellista varmistua jokaisen kohteen turvallisuudesta. Myös käyttäjien pääsynhallinta jokaiseen kohteeseen on näin vaivalloisempaa.



KUVIO 1 Massadatan automaattinen jakelu (Auto tiering)

Massadatan varsatoimisen haavoittuvuutta ja tietoturvallisuuskriittisyyttä vahvistaa myös se, että suuret datavarannot houkuttelevat helposti hyökkääjiä (Antikainen et al., 2016). Nämä seikat yhdessä johtavat massadatavarantojen datavuodon kohonneeseen riskiin. Tästä syystä massadatan säilömiseen käytetyn infrastruktuurin turvallisuudesta on tärkeä huolehtia.

Datan laatu vaikuttaa oleellisesti siitä saatavaan hyötyyn. Heikkolaatuinen data haaskaa datan varastoimiseen ja siirtämiseen vaadittuja resursseja, ja datan laatua voivat rajoittaa monet tekijät. Datan luominen, kerääminen, siirtäminen ja analysointi ovat kaikki vaiheita, joissa datan laatu saattaa kärsiä. Laadukkaan datan ominaisuuksina pidetään muun muassa tarkkuutta, kokonaisuutta, ja johdonmukaisuutta. (Chen, Mao & Liu, 2014.) Massadatan laadun puutteellisuus nousee kyber- ja tietoturvallisuuden näkökulmasta tarkasteltuna keskeiseksi siinä vaiheessa, kun massadataa hyödynnetään turvallisuuskriittisissä toiminnoissa. Tässä tilanteessa on erityisen tärkeää varmistua, että kerätty data on tarkoituksenmukaista, ja analysointiin käytetyt työkalut ja osaaminen ovat riittävällä tasolla. On esimerkiksi mahdollista, että massadata-analytiikkaa hyödyntävä tekoäly tai koneoppimisalgortimi suorittaa teollisuusympäristössä tai muussa kriittisessä infrastruktuurissa turvallisuuskriittistä toimenpidettä, ja hyökkääjä johtaa tekoälyn suorittamaan toimenpidettä epäoptimaalisesti syöttämällä sille väärennettyä dataa oikean datan seassa (CSA, 2013). Massadatan laadun varmistaminen ja analysointimenetelmien oikeaoppisuus on keskeinen haaste tekoälyn kanssa toimiessa. Tämä konkretisoituu myös esimerkiksi automaattisesti ohjautuvien autojen tapauksessa. Huolimattomuus johtaa pahimmassa tapauksessa vaarallisten uhkien toteutumiseen kyberfyysisessä ympäristössä.

Massadata esittää tietoturvahaasteita myös enkryptaus-, eli salausmenetelmien näkökulmasta. Aiemmat pienempiin ja yksinkertaisempiin datajoukkoihin sovelletut salausmenetelmät ovat osoittautuneet massadatan tapauksessa riittämättömiksi sen suuren määrän ja monipuolisuuden johdosta. 2015 toteutetussa tutkimuksessa todettiin (Toshniwal et al., 2015), että silloiset nopeimmatkin salausalgoritmit salasivat dataa 64,3 megatavua sekunnissa. Massadatan tapauksessa datan koko saattaa kuitenkin liikkua jopa petatavuissa, ja salausalgortimien nopeudesta muodostuu merkittävä pullonkaula. Massadatan käsittelyssä kerääminen ja tulosten tarkastelu tapahtuu usein myös reaaliajassa, ja sen rinnakkainen salaaminen muodostuu erityisen haasteelliseksi tästäkin syystä. Vaikka kerätty data onnistuttaisiinkin salaamaan, sen salaus pitää tyypillisesti myös purkaa ennen kyselyiden suorittamista. Tämä salauksen purkaminen ja kyselyiden suorittaminen vaatii massadatan tapauksessa merkittävästi aikaa ja laskentatehoa salauksen lisäksi.

Niin kutsuttu esineiden internet, eli internet of things tai lyhyesti IoT, on suuri nykyinen massadatan lähde, ja tuottaa todella suuren määrän dataa. Esineiden internet esittää monia tietoturvahaasteita, mutta massadatan kontekstissa yksi tällainen keskeinen riski on haitallisen toiminnan huomaamatta jääminen, sillä sen havaitseminen saattaa olla näin massiivisen datajoukon seasta hyvin haasteellista. Kun datanlähteitä on riittävän monta, muodostuu haasteelliseksi arvioida jokaisen lähteen ja datapaketin luotettavuus, sekä suodattaa mahdolliset haitalliset lähteet ja paketit. (CSA, 2013) Tämä pätee suuriin datamassoihin myös yleisellä tasolla, mutta esineiden internetin haavoittuvuuden ja kriittisyyden vuoksi se on niiden tapauksessa erityinen ongelma. Esineiden internet esittää tietoturvallisuuden kannalta merkittäviä riskejä, sillä sitä hyödynnetään yleensä reaalimaailman asioiden hoitamiseen. Tämä tapahtuu usein myös turvallisuuskriittisissä toiminnoissa esimerkiksi teollisuus- ja terveydenhuoltoympäristössä (Simon, 2017). Esineiden internetin tietoturvariskit ovat sen yleistyessä hyvin alan tutkijoiden ja asiantuntijoiden tiedossa, mutta niiden ennaltaehkäiseminen ja niihin varautuminen on hyvin tärkeää etenkin tulevaisuutta ajatellen.

Massadataympäristössä liikkuvat suuret datamassat aiheuttavat tietoturvariskejä myös esineiden internetin ulkopuolella. Nykyään tekoäly on yleinen työkalu tietoturvariskien havaitsemisessa. Tekoälyratkaisujen avulla pystytään esimerkiksi havaitsemaan automatisoidusti anomalioita, kuten poikkeavaa liikennettä organisaation tietoverkoissa. (Sarker et al., 2021) Konseptina tämä ei ole uusi, mutta modernien tekoälyratkaisujen avulla tämä on nykyään tarkempaa ja hienovaraisempaa. Erilaisia työkaluja ja algoritmeja hyödyntämällä pystytään siis nykyään havaitsemaan automatisoidusti hienovaraisiakin kyberhyökkäyksiä, jotka saattavat jäädä perinteisemmiltä palomuureilta huomaamatta. Massadatateknologioiden tapauksessa tämän tietoliikenteen ollessa hyvin suurikokoista, vaihtelevaa ja arvaamatonta, muodostuu poikkeamien havainnoiminen haasteellisemmaksi. Tekoäly saattaa työkaluna tehdä niin kutsuttuja ”false-positive” tai ”false-negative” havaintoja etenkin massadataympäristössä. Sen saattaa siis olla vaikea tunnistaa näin arvaamattomassa ja epäjoh-

donmukaisessa ympäristössä millainen toiminta on normaalia ja millainen ei, ja tästä syystä monilla reaaliaikaisen valvonnan työkaluilla on haasteita massadatan suhteen. (Lyamin et al., 2018)

Toisaalta tällaisen massiivisen tietoliikenteen manuaalinen tarkastelu ihmisen toteuttamana ei myöskään ole realistista. Tämän vuoksi joudutaan usein tekemään kompromisseja, ja tyytymään automaattisten työkalujen ja algoritmien epätäydellisyyteen. Kyberturvallisuudessa ja tietoturvallisuudessa on aina kyse kompromissien tekemisestä. Toiminnan turvallisuutta ei voida koskaan taata täydellisesti, ja joudutaan tyytymään riittävän tehokkaisiin ratkaisuihin, jotka ovat organisaation resurssit ja osaaminen huomioon ottaen realistisesti toteutettavissa. (Radziwill et al., 2017) Tästä syystä suuresta ja nopeasta dataliikenteestä automatisoidusti poikkeamia havainnoivien työkalujen ja tekoälyratkaisujen kehittäminen on massadatan tietoturvaongelmien kannalta kriittisessä asemassa. Organisaatioiden pitää kyetä turvaamaan tietoturvallisuutensa mahdollisimman kustannustehokkaasti yhteiskunnan kyberturvallisuuden takaamiseksi.

3.3 Massadatan tietoturvariskien hallitseminen

Tässä luvussa esitetään lyhyesti joitakin menetelmiä ja käytänteitä, joiden avulla edellä esitettyjä massadatan tietoturvahaasteita voidaan hallita, ja jotka auttavat niihin varautumisessa.

Kuten edellisessä kappaleessa esitettiin, massadatan automaattinen jakaminen palvelimien välillä on tyypillistä. Tästä puhutaan yleensä myös "tieringinä", jolla tarkoitetaan vähemmän tärkeän datan dynaamista ja automaattista siirtämistä tärkeää dataa alemmalle tasolle. Alemmalla tasolla oleva data taas säilötään eri palvelimilla tai muilla tallennusvälineillä, jotka ovat yleensä halvempia, ja usein myös huonommin suojattuja. (CSA, 2013.) Tallennusvälineiden infrastruktuurin ja ohjelmistojen kyberturvallisuuden varmistaminen on luonnollisesti tärkeää, ja tästä ei pitäisi karsia halvempienkaan välineiden tapauksessa, mikäli niitä käytetään tieringissä, jossa samaan ekosysteemiin liittyy tärkeää dataa. Tämän ongelman ratkaisemiseksi yritysten tulee suunnitella tarkkaan massadatan jakeluun liittyvä strategiansa, ettei sensitiivinen data päädy huonosti suojattuihin kohteisiin. Datan automaattisessa jakelussa on tärkeää myös sen liikkeiden ja sijainnin seuraaminen lokitiedostojen avulla. (Shucheng et al., 2010.) Salausmenetelmiä on kehitetty palvelimien välillä automaattisesti jaeltavan datan salaamiseksi, ja näitä on kannattavaa hyödyntää, kun mahdollista. Massadatan osalta suorituskyky näissä on kuitenkin edelleen haasteellinen kysymys. (Toshniwal et al., 2015.)

Massadatan käsittelyssä hyödynnetään NoSQL-tietokantoja. NoSQL-tietokannat soveltuvat tehtävään, sillä niiden rakenteellisen joustavuuden vuoksi ne ovat hyvin suorituskykyisiä ja helposti skaalautuvia. Nämä samat tekijät kuitenkin johtavat myös niiden suurimpiin turvallisuushaasteisiin. NoSQL-kehitettiin suurien datamassojen haasteita varten, mutta turvallisuuste-

kijöihin on kiinnitetty vain rajallisesti huomiota. (Okman et al., 2011.) NoSQL-tietokantaratkaisujen turvaamiseksi voidaan käyttää väliohjelmistoja. Tässä tapauksessa väliohjelmisto voi käyttäytyä ikään kuin kuorena, joka suojaa NoSQL-tietokantaa. Väliohjelmistoa hyödyntämällä tietokantaan ei siis pääse käsiksi, muuta kuin sen kautta. (CSA, 2013.)

Haitallisen toiminnan havaitsemiseen ja ehkäisemiseen massadatan seasta ei ole täydellistä ratkaisua. Etenkin laajoissa IoT-infrastruktuureissa jokaisen päätelaitteen tarkkailu manuaalisesti on käytännössä mahdotonta. Laitteet saattavat myös usein hälyttää virheellisesti toimiessaan normaalisti, ja väärien hälytysten suuren määrän johdosta jokaista poikkeamaa ei välttämättä edes tarkasteta. Organisaatioiden tulee huolehtia tehokkaista automatisoiduista menetelmistä ja algoritmeista liikenteen tarkkailussa. Näissä voidaan hyödyntää olemassa olevia tilastollisia malleja ja menetelmiä poikkeamien tunnistamiseksi. Nämä liikenteen tarkkailussa hyödynnettävät automatisoidut menetelmät perustuvat usein tekoälyyn ja koneoppimiseen. Avainasemassa tässä on myös datan hallinnassa käytettyjen tietoverkkojen, palvelinjärjestelmien ja ohjelmistojen turvallinen suunnittelu. (CSA, 2013.) Nämä samat menetelmät pätevät osaltansa massadataa hyödyntävien turvallisuuskriittisten tekoäly- ja koneoppmistoimintojen suojaamiseen.

Yksityisyydensuojan osalta yritysten on tärkeää olla tietoisia, että datan anonymisointi tai pseudonymisointi ei aina ole riittävä ratkaisu yksityisyydensuojan takaamiseksi. Datanlouhinnalla pystytään järjestämään tietoja uudelleen alkuperäisten henkilötietojen paljastamiseksi. (Strohbach, 2016) Henkilötietoja massadatassa ei välttämättä voida salata täysin varmasti, mutta tietoisuus anonymisoinnin tai pseudonymisoinnin riittämättömyydestä auttaa varautumaan mahdollisiin ongelmiin. Datan oikeaoppinen anonymisointi ja pseudonymisointi on silti tärkeää henkilötietoja käsitellessä, vaikka se ei olisikaan aina riittävä ratkaisu. Organisaatioiden tulee olla tietoisia GDPR:n säädöksistä, ja käsitellä henkilötietoja ohjeistuksen mukaisesti. Etenkin GDPR:n käyttäjälähtöisen suunnittelun periaatteesta on tärkeää olla tietoinen.

4 Pohdinta ja yhteenveto

Tässä luvussa tiivistetään massadatan hyödyntämistä, ja sen keskeisimpiä tietoturvariskejä ja -ongelmia, vastaten samalla tutkielman tutkimuskysymyksiin.

Massadata on dataa, jota on paljon, sitä tulee nopeasti lisää, ja se on luonteeltaan monipuolista ja hajanaista. Massadataa hyödynnetään nykyään liki joka alalla, ja sen hyödyntäminen on edelleen jatkuvassa nousussa. ETLAn kyselytutkimuksessa todettiin, että jo vuonna 2015 massadata oli erittäin relevantti aihe suomalaisten yritysten keskuudessa, ja nykyään sen hyödyntäminen on entistä aktiivisempaa. Myös kansainvälisissä tutkimuksissa datatalouden ja massadatateknologien nousu on nähty jo pitkään selkeästi, eikä viitteitä muutokseen lähiaikoina ole huomattavissa.

Tutkielman ensimmäiseen tutkimuskysymykseen ”Miten organisaatiot hyödyntävät massadataa toiminnassaan?” vastattiin kattavasti luvussa kaksi. Siihen voidaan tiivistää tutkielman pohjalta vastaus esimerkiksi seuraavasti: Organisaatiot keräävät massadataa monenlaisista lähteistä erilaisilla menetelmillä. Tyypillisiä lähteitä ovat esimerkiksi verkkosivut ja -palvelut, joiden kautta kerätään muun muassa niin kutsuttua ”ihmisten tuottamaa dataa”. Tätä dataa kerätään esimerkiksi lokitiedostojen, evästeiden ja ”web crawlerien”, eli hakurobottien avulla. Massadataa kerätään myös fyysisestä ympäristöstämme erilaisilla sensoreilla. Tämä on erittäin yleistä nykyään IoT-, eli esineiden internet -ratkaisujen yleistyessä. IoT on yleisesti käytössä myös teollisuusympäristössä, usein turvallisuuskriittisissä infrastruktuureissa. Organisaatiot keräävät näistä lähteistä massiivisia määriä raakadataa, jota säilötään ja käsitellään tietovarannoissa, jotka sijoittuvat usein monelle eri palvelimelle. Tästä raakadatasta jalostetaan tiedonlouhinnan ja data-analytiikan menetelmillä organisaatioille hyödyllistä informaatiota. Tätä informaatiota voidaan käyttää esimerkiksi markkinoinnin kohdentamiseen oikeanlaiselle yleisölle, tai julkisen liikenteen suunnitteluun. Sovelluskohteita massadatalle on rajattomasti, mutta kyberturvallisuuden kannalta merkittävimpiä ovat turvallisuuskriittiset sovelluskohteet, kuten kriittisissä toiminnallisuuksissa hyödynnettävät koneoppimisalgoritmit.

Massadatan hyödyntäminen tarjoaa organisaatioille monia mahdollisuuksia ja etuja, mutta se esittää myös haasteita tietoturvallisuuden näkökulmasta.

Yksi tähän liittyvistä keskeisimmistä haasteista on yksityisyydensuojan säilyttäminen massadataa käsitellessä. GDPR:n mukaan anonymisoituja henkilötietoja voidaan käsitellä vapaasti, mutta data-analytiikan ja tiedonlouhinnan menetelmillä on mahdollista paljastaa anonymisoiduista tai pseudonymisoiduista tiedoista aiemmat henkilötiedot. Organisaatioiden on hyvä olla tietoisia tästä ja soveltaa salausmenetelmiä aina kun mahdollista.

Massadatan hyödyntämiseen sovelletaan tyypillisesti NoSQL-tietokantoja suorituskyvyn varmistamiseksi, mutta NoSQL-tietokannoissa on monia puutteita tietoturvallisuudessa, joita normaaleissa relaatiotietokannoissa ei ole. NoSQL-tietokantoja voidaan kuitenkin upottaa hyvin suojattuihin väliohjelmistoihin, jotka käyttäytyvät niille suojakuorena. Massadatan ollessa usein säilöttynä klustereina monella eri palvelimella, organisaatioiden saattaa olla haastava huolehtia jokaisen palvelimen ja tallennusmedian turvallisuudesta. Näiden datan säilömiseen liittyvien haasteiden lisäksi organisaatioiden tulee olla huolellisia datan automaattisen jakelun suunnittelussa palvelimien välillä. Tämän turvallisuutta voidaan edesauttaa lokitiedostoja hyödyntämällä ja hyvin suunnitellulla jakelustrategialla. Myös salausalgoritmeja voidaan hyödyntää massadatan käsittelyssä ja jakelussa, mutta näiden suorituskyky massadatan suhteen on edelleen ongelmallinen kysymys. Tässä ollaan kuitenkin tultu eteenpäin viime vuosina. Massadatavarantojen turvallisuuden suunnittelu on tärkeää myös siitä syystä, että suuret datavarannot houkuttelevat hyökkääjiä.

Massadatan aiheuttaman liikenteen ollessa tyypiltään suurta, nopeaa ja hajanaista, saattaa haitallisen toiminnan havaitseminen tämän seasta olla haasteellista. Organisaatioiden tulee käyttää mahdollisimman tehokkaita automaattisia koneoppimisalgoritmeja ja tekoälyratkaisuja haitallisen liikenteen tunnistamiseksi. Nämä pohjautuvat tilastollisiin malleihin poikkeamien tunnistamiseksi. Tässä on luonnollisesti keskeistä myös dataa jakelevien ja käsittelevien palvelinjärjestelmien ja -ohjelmistojen turvallinen suunnittelu. Massadatan automaattisen jakelun tapauksessa myös alemman tason palvelimien ja tallennusmedioiden turvallisuuden varmistaminen on tärkeää.

Tutkielman toiseen tutkimuskysymykseen ”Millaisia tietoturvaongelmia massadatan hyödyntäminen aiheuttaa?” vastattiin kattavasti luvussa kolme. Siihen voidaan tiivistää tutkielman pohjalta vastaus seuraavasti:

- Henkilötietosuojan säilyttäminen anonymisoinnin ja pseudonymisoinnin ollessa keinoina ajoittain riittämättömiä.
- NoSQL-tietokantojen puutteellinen tietoturvallisuus.
- Datan automaattisen palvelinjakelun tietoturvahaasteet.
- Salausalgoritmien suorituskyvyn haasteet suuria datamassoja käsiteltäessä.
- Haitallisen toiminnan ja liikenteen havaitseminen massadatan seasta, esimerkiksi laajoissa IoT-infrastruktuureissa.
- Massadatan oikeaoppinen analysointi ja käsittely turvallisuuskriittisiä toimintoja toteutettaessa. Tämä konkretisoituu esimerkiksi tekoälyn toiminnassa automaattisesti ohjautuvissa autoissa.

Massadatan kaltaiset teknologiat halutaan usein nähdä vain niiden tuomien liiketoimintaetujen näkökulmasta, eikä niiden kokonaisvaltaisuutta aina nähdä riittävän kattavasti. Organisaatioiden on tärkeää olla tietoisia myös hyödyntämiensä teknologioiden tieto- ja kyberturvallisuudesta. Massadata esittää monia tietoturvaongelmia, mutta nämä ovat kuitenkin hallittavissa datanhallinnan oikeaoppisella suunnittelulla. Tutkielman pohjalta harkitsemisen arvoisia jatkotutkimusaiheita ovat esimerkiksi:

- Datan anonymisoinnin haasteet ja kehityskohteet massadatan kontekstissa.
- Haitallisen toiminnan tunnistaminen massadatasta tekoälyn avulla.
- Käytänteet massadatan tietoturvahaasteisiin varautumiseksi.
- Massadatan käsittelyyn soveltuvien salausalgoritmien kehitys.

LÄHTEET

A. Labrinidis and H. Jagadish, "Challenges and opportunities with big data," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2032–2033, 2012.

A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," *Journal of Biomedical Informatics*, vol. 46, no. 5, pp. 774–781, 2013.

Acquisti, A., & Gross, R. (2009). Predicting social security numbers from public data. Proceedings of the National Academy of Sciences, 106(27), 10975–10980

Boyd, Dana; Crawford, Kate (21 September 2011). "*Six Provocations for Big Data*". Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society.

Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." *Mobile networks and applications* 19.2 (2014): 171-209.

Computer Sciences Corporation. (2014). Big data Universe beginning to explode. http://www.csc.com/insights/flxwd/78931-big_data_universe_beginning_to_explode

CSA (Cloud Security Alliance), Expanded Top Ten Big Data Security and Privacy Challenges. CSA Report, 16 June 2013. https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

Demchenko, Y., Gruengard, E. & Klous, S., 2014. Instructional Model for Building Effective Big Data Curricula for Online and Campus Education. 2014 IEEE 6th International Conference on Cloud Computing Technology and Science, pp.935–941. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7037787

Dongpo Zhang (2018). Big Data Security and Privacy Protection. Advances in Computer Science Research, volume 77

Erevelles, Sunil, Nobuyuki Fukawa, and Linda Swayne. "Big Data consumer analytics and the transformation of marketing." *Journal of business research* 69.2 (2016): 897-904.

e-skills UK, 2014. Big Data Analytics: Adoption and Employment Trends, 2012–2017.

http://www.sas.com/offices/europe/uk/downloads/bigdata/eskills/eskills.pdf

General Data Protection Regulation (GDPR) – Official Legal Text, 2018: https://gdpr-info.eu/

Gudivada, Venkat N., Dhana Rao, and Vijay V. Raghavan. "NoSQL systems for big data management." *2014 IEEE World congress on services*. IEEE, 2014.

H. Cai, B. Xu, L. Jiang and A. V. Vasilakos, "IoT-Based Big Data Storage Systems in Cloud Computing: Perspectives and Challenges," in *IEEE Internet of Things Journal*, vol. 4, no. 1, pp. 75-87, Feb. 2017, doi: 10.1109/JIOT.2016.2619369.

Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies*. John Wiley & Sons.

IDC (2015). European Data Market SMART 2013/0063. D6- First Interim Report. Oct 16th 2015.

IDG Enterprise (2015). Big Data and Analytics Survey. http://www.idgenterprise.com/report/2015-big-data-and-analytics-survey

J. Fan and H. Liu, "Statistical analysis of big data on pharmacogenomics," *Advanced Drug Delivery Reviews*, vol. 65, no. 7, pp. 987–1000, 2013.

Janne Antikainen, Jarmo Eskelinen, Heli Koski, Tommi Niemi, Mika Pajarinen, Sinikukka Pyykkönen, Marc de Vries. 2016. Valtioneuvoston selvitys-ja tutkimustoiminnanjulkaisusarja 16/2016. Massadatasta liiketoimintaa ja tehokkaita julkisia palveluja. ETLA.

Jonker, W. & Petković, M. (2012). Secure Data Management: 9th VLDB Workshop, SDM 2012, Istanbul, Turkey, August 27, 2012: Preface. *Lecture Notes in Computer Science, 7482*, urn:issn:0302-9743.

K. Douglas, "Infographic: big data brings marketing big numbers", 2012, http://www.marketingtechblog.com/ibm-big-data-marketing/.

K. Michael and K. W. Miller, "Big data: new opportunities and new challenges," *Editorial: IEEE Computer*, vol. 46, no. 6, pp. 22–24, 2013.

Liang, Fan, et al. "A survey on big data market: Pricing, trading and protection." *IEEE Access* 6 (2018): 15132-15154.

Lundgren, Björn, and Niklas Möller. "Defining information security." *Science and engineering ethics* 25.2 (2019): 419-441.

Lyamin, Nikita, et al. "AI-based malicious network traffic detection in VANETs." *IEEE Network* 32.6 (2018): 15-21.

M. Chen, S. Mao, and Y. Liu, "Big data: a survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171–209, 2014.

Madakam, S. , Ramaswamy, R. and Tripathi, S. (2015) Internet of Things (IoT): A Literature Review. *Journal of Computer and Communications*, **3**, 164-173. doi: 10.4236/jcc.2015.35021.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., (2011) Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

Mayer-Schönberger, Viktor – Cukier, Kenneth (2013). Big Data – A Revolution That Will Transform How We Live, Work and Think. John Murray.

Miranda Mourby, Elaine Mackey, Mark Elliot, Heather Gowans, Susan E. Wallace, Jessica Bell, Hannah Smith, Stergios Aidinlis, Jane Kaye, Are 'pseudonymised' data always personal data? Implications of the GDPR for administrative data research in the UK, Computer Law & Security Review, Volume 34, Issue 2, 2018, Pages 222-233, ISSN 0267-3649

N. Chaudhari and S. Srivastava, "Big data security issues and challenges," *2016 International Conference on Computing, Communication and Automation (ICCCA)*, Noida, 2016, pp. 60-64, doi: 10.1109/CCAA.2016.7813690.

Nathan L., Nicola H., Roger G. & Kiersten M. "Please prove the claim that 90 % of the world's data has been generated in the last two years, including hard statistics to back it up", Wonder, 2017 https://askwonder.com/research/please-prove-claim-90-world-s-data-generated-past-two-years-including-hard-utp5zijln

Nawsher Khan, Ibrar Yaqoob, Ibrahim Abaker, Targio Hashem, Zakira Inayat, Waleed Kamaleldin Mahmoud Ali, Muhammad Alam, Muhammad Shiraz, Abdullah Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges", *The Scientific World Journal*, vol. 2014, Article ID 712826, 18 pages, 2014. https://doi.org/10.1155/2014/712826

Okman, L., Gal-Oz, N., Gonen, Y., Gudes, E., & Abramov, J. (2011, November). Security issues in nosql databases. In *2011IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications* (pp. 541-547). IEEE.

P. Russom, "Big data analytics," *TDWI Best Practices Report*, Fourth Quarter, 2011.

Radziwill, Nicole M., and Morgan C. Benton. "Cybersecurity cost of quality: Managing the costs of cybersecurity risk management." *arXiv preprint arXiv:1707.02653* (2017).

Raghav Toshniwal, Kanishka Ghosh Dastidar, Asoke Nath (2015). Big data security issues and challenges. International Journal of Innovative Research in Advanced Engineering (IJIRAE). ISSN: 2349-2163. Issue 2, Volume 2

Sarker, Iqbal H., Md Hasan Furhad, and Raza Nowrozy. "Ai-driven cybersecurity: an overview, security intelligence modeling and research directions." *SN Computer Science* 2.3 (2021): 1-18.

Shucheng, Y., Wang, C., Ren, K., & Lou, W. (2010). Achieving secure, scalable, and fine-grained data access control in cloud computing. INFOCOM (pp. 1–9).

Simon, T. (2017). Chapter seven: Critical infrastructure and the internet of things. *Cyber Security in a Volatile World*, *93*.

Strohbach, M., Daubert, J., Ravkin, H., & Lischka, M. (2016). Big data storage. In *New horizons for a data-driven economy* (pp. 119-141). Springer, Cham.

Toshniwal, Raghav, Kanishka Ghosh Dastidar, and Asoke Nath. "Big data security issues and challenges." *Complexity* 2.2 (2015).

Winder, D. (2012). Securing NoSQL applications: Best practises for big data security. Computer Weekly

Wolkowitz, Eva – Parker, Sarah (2015). Big Data, Big Potential: Harnessing Data Technology for the Underserved Market. Center for Financial Services Innovation.

Wu, Xindong, et al. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26.1 (2013): 97-107.