Henri Talviaho

NÄKEMYKSIÄ BIG DATASTA: PALJON DATAA, VAI JOTAIN ENEMMÄN?

JYVÄSKYLÄN YLIOPISTO
INFORMAATIOTEKNOLOGIAN TIEDEKUNTA
2021

TIIVISTELMÄ

Talviaho, Henri
Näkemyksiä Big Datasta: Paljon dataa, vai jotain enemmän
Jyväskylä: Jyväskylän yliopisto, 2021, 61 s.
Tietojärjestelmätiede, pro-gradu tutkielma
Ohjaaja: Taipalus, Toni

Datan määrä on kasvanut räjähdysmäisesti. Datan rooli yhteiskunnassa on muuttunut 2000-luvulla merkittävästi. Dataa hyödynnetään monilla eri tavoilla, esimerkiksi markkinoinnissa. Big Datalla tarkoitetaan dataa, joka on määrältään suurta ja olemukseltaan moninaista. Big Datalle ei kuitenkaan ole olemassa yksittäistä käsitettä, vaan niin yritysmaailma, kuin myös akateeminen maailma ovat pullollaan useita toisistaan poikkeavia käsitteitä. Tässä tutkielmassa tavoitteena on kartoittaa niin kirjallisuuskatsauksen kuin myös empiirisen tutkimuksen avulla sitä, mitä Big Datalla tarkoitetaan ja sitä, minkälaisia väärinkäsityksiä Big Dataan liittyy. Tutkielmassa tutustutaan aluksi kuvailemaan sitä, miten akateemikot näkevät Big Datan käsitteen tasolla. Lisäksi tutustutaan Big Datan yleisimpiin ominaisuuksiin. Empiirisen tutkimuksen avulla tutustutaan siihen, miten asiantuntijoiden ja opiskelijoiden piirissä Big Data mielletään. Tutkimuksessa havaittiin, että Big Dataan liittyviä väärinkäsityksiä ovat yleisimmin se, että Big Datan koetaan olevan vain määrällisesti suurta dataa ja täten eroavan vain määrän perusteella tavallisesti datasta. Lisäksi toisena yleisimpänä väärinkäsityksen muotona tutkimuksessa ilmeni se, että Big Data sekoitetaan analytiikan kanssa, toisin sanoen vastaajat olettavat, että Big Data itsessään on prosessi, joka kattaa kaiken tiedon keräämisestä analysointiin saakka. Tutkimuksen lopussa pohditaan Big Datan tärkeyttä ja tarpeellisuutta. Kyseenalaistetaan termin Big Data tarpeellisuus ja ehdotetaan, että Big Data olisi vain osa data-analytiikkaa.

Avainsanat: Big Data, Big Data analytiikka, data-analytiikka, data, small data

ABSTRACT

Talviaho, Henri
Views on Big Data: Large on volume or something more?
Jyväskylä: University of Jyväskylä, 2021
Information Systems, Master's thesis
Supervisor: Taipalus, Toni

The amount of data has grown significantly. Nowadays data is collected in many different ways and with different means. Data is, for example, used to profile us and in marketing. Big Data has multiple different definitions. Usually Big Data is seen as data that is huge in volume and rich in form. Most of academics explain the Big Data with three feature, volume, velocity and variety. Because of the lack of accepted definition, defining Big Data is challenging. The aim of this Master's thesis is to investigate how Big Data is defined in academia and in practice, and what misunderstanding and misconceptions of Big Data excist. The study found that the most common misconceptions about Big Data are that Big Data is perceived to be only quantitatively large data and thus differs only with volume when compared to basic data. the second most common form of misunderstanding in the study was that Big Data is confused with analytics, i.e., respondents assume that Big Data itself is a process that covers everything from data collection to analysis. At the end of the study, the importance and necessity of Big Data is considered. The necessity of the term Big Data is questioned and it is suggested that Big Data should only be part of data analytics.

Keywords: Big Data, Big Data Analytics, data-analytics, data, analytics, small data

KUVIOT

TAULUKOT

SISÄLLYS

1 JOHDANTO

Datan määrä on kasvanut räjähdysmäisesti. Eatonin ym. (2012) mukaan datajätit Facebook ja Twitter generoivat dataa useita teratavuja päivittäin. Suuri datan määrä aiheuttaa haasteita eri instansseille hyödyntää sitä. Erilaisia menetelmiä Big Datan hyödyntämiseen onkin kehitetty.

Big Datalle ei kuitenkaan ole olemassa tarkkaa määritelmää. Big Data terminä ei kuitenkaan ole kovinkaan uusi, sillä Cox ja Ellsworth (1997) puhuivat omassa tutkimuksessaan Big Datasta jo vuonna 1997. Ehkä ensimmäinen laajalti hyväksyntää saanut määritelmä Big Datasta on Doug Laneyn vuonna 2001 tekemä yleistys Big Datasta (Kitchin & McArdle, 2016). Hänen mukaansa Big Dataa voidaan kuvata kolmella keskeisellä ominaisuudella, jotka ovat datan määrä, nopeus ja moninaisuus. Ajan kanssa on kuitenkin kehitetty lisää kuvaavia ominaisuuksia Big Datalle. On kuitenkin syytä kyseenalaistaa kaikkien kuvaavien ominaisuuksien tarpeellisuus.

On myös olemassa muita tapoja määrittää Big Dataa, sillä Big Dataa kuvaavia käsitteitä on vuosien varrella kehitetty samankaltaisia määriä kuin ominaisuuksiakin. Osa tutkijoista näkee Big Datan pelkästään siihen liittyvien ominaisuuksien, kuten merkittävän datamäärän, pohjalta. On kuitenkin olemassa myös toisenlaisia tulkintoja, jotka eivät pelkästään pohjaudu Big Dataa kuvaaviin ominaisuuksiin.

Big Dataa voidaan hyödyntää monilla eri tavoin monilla eri aloilla. Rahoituksessa ja finanssialalla Big Dataa hyödynnetään monissa eri käyttötarkoituksissa, esimerkiksi ennustamaan ja laskemaan asiakkaan lainanmaksukykyä (Hussain & Prieto, 2016), pankkipalveluiden optimoinnissa (Bedeley, 2014) ja apuna tunnistamaan mahdollisia huijauksia rahoitusalalla (Sharma, Pandey & Kumar, 2016). Big Dataa hyödynnetään merkittävästi myös muilla aloilla. Esimerkiksi terveysalalla Big Datan avulla pystytään ennustamaan esimerkiksi flunssakausia, pandemioita ja optimoimaan tehohoidon kapasiteettia (Bates ym., 2014; Andreu-Perez ym., 2015; Feldman, Martin & Skotnes, 2012). Big Dataa voidaan hyödyntää myös muilla aloilla, esimerkiksi matkailussa (Chen ym., 2016) hyödyntämällä dataa esimerkiksi ihmisten liikkumisen ennustamisessa.

Big Datan hyödyt ovat selvät. Mutta onko Big Data enemmänkin merkittävää suosiota osakseen saanut ilmiö, vai jotain merkittävämpää? Viimeisten vuosien aikana Big Datan suosio on kasvanut merkittävästi, datan määrän mukana. Sitä mukaa kun Big Datan suosio on kasvanut, on myös erilaisia määritelmiä Big Datalle kehitetty samaan tahtiin. Teknologisten harppausten ja kehitettyjen ohjelmistojen avulla Big Datan kuvailemisesta on tullut aiempaa haastavampaa. Enää ei riitä Big Datan käsitteeksi se, että dataa on paljon. Mikä itseasiassa sitten on paljon, riippuu kontekstista. Useammin Big Dataa voitaisi kuvailla useilla ominaispiirteillä, jotka esiintyvät yhdessä.

Kaupalliset osapuolet ovat osasyyllisiä siihen, että erilaisia näkemyksiä Big Datalle on niin monia. Big Dataa on käytetty enenemissä määrin väärin vain edistämään kaupallisten tuotteiden menekkiä, unohtaen Big Datan tarkoituksen. Big Dataa kuvailemaan on kehitetty entistä villimpiä ominaisuuksia ja ominaispiirteitä kuvaamaan sitä.

Big Datan hyödyt ovat varmasti selvät. On kuitenkin kyseenalaistettava ihmisten tietämys Big Datasta. Termiin Big Data liittyy tietynlaista mystisyyttä ja mystiikkaa. Monelle Big Data on edelleen vain dataa, mutta monille se on myöskin paljon enemmän. On kuitenkin selvää, ettei ole olemassa yhtä oikeaa vastausta siihen, mitä Big Data on. Ennemminkin maailma on pullollaan erilaisia näkemyksiä, osa oikeanlaisia, osa vääriä ja osa jopa harhaanjohtavia. On kuitenkin selvää, että Big Data on tullut jäädäkseen, mutta on myös selvää, että se tarvitsee osakseen käytännönläheisemmän käsitteen.

1.1 Kirjallisuuskatsaus

Tämä tutkielma koostuu kahdesta pääosasta, kirjallisuuskatsauksesta ja empiirisestä osuudesta. Kirjallisuuskatsauksen tavoitteena on ollut tarjota Big Datasta mahdollisimman hyvä kuvaus siten, ettei lukijalla ole tarvetta etsiä lisää tietoa muualta.

Kirjallisuuskatsauksen työstämisessä on hyödynnetty Templierin ja Parén (2015) luomaa viitekehystä, joka koostuu kuudesta pääpiirteestä. Nämä piirteet ovat:

1) ongelman määrittäminen

2) kirjallisuuden etsintä

3) seulonta sisällyttämistä varten

4) laadun arviointi

5) tiedon louhinta

6) kerätyn tiedon analysointi ja yhdistely

Ongelman määrittäminen pitää sisällään tutkimusasetelman asettamisen, tärkeimpien käsitteiden tunnistamisen sekä perustelun sille, miksi kirjallisuuskatsaus pitäisi tehdä. Kirjallisuuden etsintä keskittyy nimensä mukaisesti kirjalli-

suuden kartoittamiseen. Etsinnän lisäksi tässä vaiheessa tavoitteena on tunnistaa hyödylliset lähteet tiedolle sekä yksilöitä tutkimuksia, jotka ovat hyödyllisiä katsauksen kannalta. Seulonta sisällyttämistä varten tarkoittaa vaihetta, jossa tutustutaan kerättyyn aineistoon. Arvioidaan sen soveltuvuutta asetettuun tutkimusongelmaan. Vaihe sisältää kirjallisuuden valintaa ja pois sulkemista. Laadun arviointi keskittyy arvioimaan valitun kirjallisuuden laadukkuutta. Laadukkuutta voidaan arvioida esimerkiksi julkaisufoorumin avulla, joka pisteyttää halutun lähteen julkaisualustan perusteella. Tiedon louhinnassa tavoitteena on kerätä tietoa valitusta lähdekirjallisuudesta ja viimeinen vaihe, kerätyn tiedon analysointi ja yhdistely, keskittyy aineiston järjestelyyn, vertailuun ja tiivistykseen (Templier & Paré, 2015).

Kirjallisuuskatsauksen lähteiden etsintään hyödynnettiin verkosta löytyviä akateemisen kirjallisuuden hakupalveluita, kuten Google Scholar, Web of science, JYKDOK, Ieee Xplore ja Scopus. Tietoa Big Datasta etsittiin näistä edellä mainituista hakupalveluista erilaisine hakulausekkeineen. Esimerkkejä hakulauseista ovat: ”Big Data features” ”Big Data analytics” ”Big Data definition” ”Big Data misconceptions” ja ”Big Data”.

Kirjallisuutta löydettiin runsaasti, mutta lähdekirjallisuuden osalta haluttiin tehdä rajaus, jonka perusteella tiettyä vuotta ennen julkaistut julkaisut jätettiin pois lähdeaineistosta. Vuodeksi valittiin 2012, mutta rajauksessa joustettiin tarvittaessa, jos havaittiin, että jollakin tietyllä julkaisulla oli tutkielman kannalta merkittävä asema. Rajaukseen päädyttiin siitä syystä, että Big Data ilmiönä on ollut viimeisten vuosien aikana suuresti pinnalla sekä jatkuvat teknologiset harppaukset ovat muuttaneet esimerkiksi näkemystä siitä, mikä määrä dataa voisi olla Big Dataa.

1.2 Tutkimusongelma ja tutkimuskysymykset

Kuten todettua, dataa syntyy nykyään merkittäviä määriä joka hetki. Dataanalytiikkaa on hyödynnetty jo pitkään esimerkiksi markkinoinnissa (Xu, Frankwick & Ramirez, 2016). Myös Big Dataa on tutkittu laajasti monista eri lähtökohdista ja näkökulmista. Big Data on itsessään saanut suurta hypetystä eri osapuolilta ja voidaan sanoa, että termin kuulemiselta ei voi välttyä.

Big Datalle ei ole olemassa tarkkaa määritelmää, vaan määritelmät vaihtelevat näkökulmien ja eri osapuolien perusteella. Monessa tapauksessa Big Dataa pyritään määrittämään kaupallisesta näkökulmasta niin, että oman yrityksen hyöty olisi mahdollisimman suurta. Tämä on johtanut siihen, että Big Dataa kuvataan mitä mielenkiintoisemmilla ominaisuuksilla ja ominaispiirteillä.

Ymmärretäänkö Big Dataa oikeasti? Mitä se tarkoittaa? Jokaisella on varmasti olemassa omanlaisensa määritelmä Big Datalle, mutta mikä määritelmistä on oikea tai oikeanlainen? Big Datan tutkimuksessa on keskitytty Big Datan määrittämiseen teoreettisesta näkökulmasta, eikä Big Dataan liittyvästä käsittämisestä ole tehty merkittävää määrää tutkimusta. Yksi ainoista tutkimuksista

tähän liittyen on Favaretton ym. (2020) tekemä tutkimus, jossa tutkittiin sitä, miten akateemiset tutkijat määrittävät Big Datan.

Tässä tutkielmassa tavoitteena on tutkia sitä, miten Big Dataa ymmärretään ja käsitetään. Voidaanko Big Datan ymmärtämisen suhteen tunnistaa yleisimpi väärinkäsityksiä ja ymmärryksiä? Big Dataan liittyvästä väärinymmärryksestä ja -käsityksestä ei ole tehty merkittävästi tutkimusta, mutta aihe itsessään on tunnistettu akateemisen kirjallisuuden ulkopuolella. Tutkimus aiheesta on mielestäni tärkeää siksi, että se motivoi muodostamaan Big Datan kannalta sellaisen käsitteen, joka on helpompi ymmärtää ja käytännönläheisempi. Lisäksi tutkimus on tärkeässä asemassa sen vuoksi, että se pystyy vahvistamaan tai hylkäämään kirjallisuudessa esitettyjä näkemyksiä.

Tutkimuskysymyksiksi on asetettu seuraavat kysymykset:

1) **Mitä Big Data tarkoittaa?**
2) **Miten Big Data koetaan käytännössä?**
3) **Liittyykö Big Dataan väärinymmärryksiä ja -käsityksiä?**

Asetetuista tutkimuskysymyksistä vastataan kirjallisuuskatsauksen perusteella ensimmäiseen kysymykseen, empiirisen osuuden avulla vastataan toiseen asetettuun kysymykseen ja lopulta näiden kahden kysymyksen vastauksien avulla pyritään vastaamaan kolmanteen asetettuun kysymykseen.

1.3 Tutkimuksen rakenne

Tutkielma koostuu seitsemästä pääluvusta ja alaluvuista. Ensimmäiset kaksi lukua johdannon jälkeen on pyhitetty kirjallisuuskatsaukselle. Toisessa luvussa keskitytään Big Dataan käsitteenä ja tunnistetaan sen kannalta olennaisia käsitteitä. Kolmannessa luvussa keksityttään analysoimaan kirjallisuutta ja löytämään valitun kirjallisuusaineiston perusteella yleisimpiä ominaisuuksia, joita kirjallisuudessa on liitetty Big Dataan. Neljännessä luvussa käydään läpi empiirisen osuuden metodologiaa, valittua tutkimusmenetelmää, esitellään tutkimuksen tausta ja valittu aineiston analysointimenetelmä. Viidennessä luvussa keskitytään tutkimuksen tuloksiin ja lopussa pyritään vastaamaan valittuihin tutkimuskysymyksiin. Kuudennessa luvussa pohditaan tutkimuksen tulosta ja sitä, mitä vaikutuksia tuloksella Big Datan kannalta on. Lopuksi seitsemännessä luvussa vedetään tutkielma yhteen, pohditaan tutkimuksen luotettavuutta ja toistettavuutta sekä pohditaan mahdollisia jatkotutkimusaiheita. Tutkielman lopussa on luettelo tutkielmassa käytetyistä lähteistä.

2 BIG DATA KÄSITTEENÄ

Tässä luvussa keskitytään Big Dataan käsitteenä. Aluksi perehdytään kirjallisuudessa esitettyihin näkemyksiin Big Datasta, jonka jälkeen keskitytään valaisemaan Big Datan eroavaisuutta small dataan. Lopuksi erityisesti Big Datan hyödyntämisen kannalta olennainen käsite, Big Data analytiikka ja sen yhteys Big Dataan käsitellään.

2.1 Big Datan luokittelu

Big Datan määrittely on haastavaa. Termille Big Data ei ole olemassa yhteistä hyväksyttyä määritelmää. Yleisimmin Big Dataa kuvaillaan kolmen V:n avulla, jotka keskittyvät kuvailemaan Big Datan ominaispiirteitä, määrää (volume), nopeutta (velocity) ja moninaisuutta (variety) (Kitchin & McArdle, 2016). Kyseiset kolme ominaisuutta ovat peräisin vuodelta 2001 Gartnerin analyytikko Doug Laneyn blogikirjoituksesta. Sittemmin eri tahot ovat laajentaneet Big Dataa kuvailevia ominaisuuksia esimerkiksi arvolla (value), vaihtelevuudella (variablity), visuaalisuudella, todenmukaisuudella (veracity) ja kompleksisuudella. Kuten todettua, Big Datan määrittely on haastavaa ja haastavuudesta oiva osoitus on se, että tutkimuksissa saatetaan puhua nykyään jopa seitsemästätoista Big Dataa kuvaavasta ominaisuudesta. Yleisimmin toistuvia ominaisuuksia Big Datan kuvaamisessa ovat kuitenkin edelleen Doug Laneyn vuonna 2001 kuvaavat määrä, vauhti ja moninaisuus.

Tutkimuksissa on yleensä keskitytty kuvaamaan Big Dataa teknologisesta näkökulmasta, erityisesti siitä, minkälaista haastetta räjähdysmäisesti kasvava datamäärä aiheuttaa nykyteknologialle. On kuitenkin myös olemassa vaihtoehtoisia tapoja määrittää Big Dataa. Esimerkiksi De Mauro, Greco ja Grimaldi (2015) kuvaavat Big Dataa eri tavalla. Heidän mukaansa Big Dataa kuvaavia teemoja ovat informaatio, teknologia, keinot ja vaikutus.

Informaatiolla De Mauro ym. (2015) tarkoittavat laajaa datan tuottamista, jakamista ja käyttöä. Heidän mukaansa yhtenä merkittävimpänä syynä Big Da-

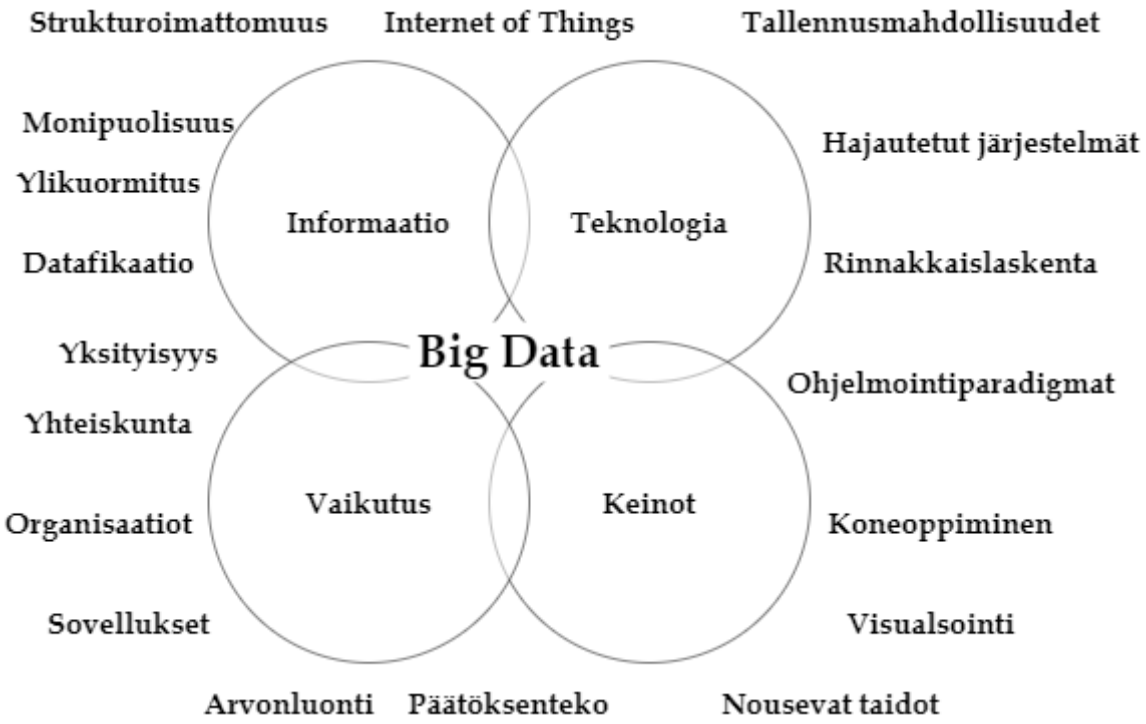
tan suosion kasvulle voidaan pitää datafikaatioita. Datafikaatiolla tarkoitetaan Southertonin (2020) mukaan prosessia, missä subjektit, objektit, prosessit ja käytännöt muutetaan digitaaliseksi tiedoksi, eli dataksi. Mayer-Schönbergerin ja Cukierin (2013) mukaan datafikaatio mahdollistaa sellaisten uusien ja uniikkien trendien ja mallien löytämisen, joiden löytämistä aiemmin on saatettu pitää jopa täysin mahdottomana, varsinkin silloin kun data on ollut analogisessa muodossa.

Toinen syy informaation roolin tärkeydessä on De Mauron ym. (2015) mukaan uudet henkilökohtaiset älylaitteet, jotka ovat täynnä erilasia sensoreita, jotka keräävät meistä tietoa jatkuvalla syötöllä. Tällaiset sensorit mahdollistavat De Mauron ym. (2015) mukaan digitalisaation samalla kun verkkoyhteys mahdollistaa datan keräämisen, muuntamisen ja lopulta myös organisoinnin tiedoksi. Gartnerin (2021) mukaan vuonna 2020 maailmassa olisi arviolta noin 26 miljardia laitetta.

Toinen Big Dataa kuvaavista teemoista on teknologia. De Mauron ym. (2015) mukaan teknologia on välttämätön esivaatimus Big Datan hyödyntämiselle. Heidän mukaansa on selvää, ettei nykyaikaiset analysointimenetelmät ole riittäviä hyödyntääkseen Big Dataa. De Mauro ym. (2015) mainitsevat esimerkkeinä Big Datan käsittelyyn soveltuvista menetelmistä esimerkiksi Apache Hadoop -viitekehyksen. Yleisimmin De Mauro ym. (2015) kuvaavat teknologialle yleisen tason vaatimuksia, jotta ne suoriutuisivat Big Datan analysoinnista. Näitä vaatimuksia on esimerkiksi prosessointikyvykkyys, kyvykkyys siirtää suuria määriä dataa ja tarpeeksi suuri kapasiteetti säilöä kerättyä dataa.

Keinot -näkökulmalla De Mauro ym. (2015) tarkoittavat niitä keinoja, joiden avulla merkittävää datamäärää voidaan käsitellä. Erilaisia keinoja hyödyntää ja käsitellä Big Dataa on useita, esimerkiksi neuroverkot, koneoppiminen, visualisointi ja regressiomallit ovat keinoja, joissa Big Dataa voidaan hyödyntää. De Mauro ym. (2015) huomattavat Big Datan kehityksen muuttaneen päätöksenteon aiemmasta staattisesta prosessista enemmän dynaamiseksi prosessiksi. Heidän mukaansa erilaiset johdannaiset datasta ovat korvanneet aiemmat tavoitteelliset loogiset yhteydet. De Mauron ym. (2015) mielestä yritysten ja organisaatioiden tulisi panostaa kriittisiin analyyttisiin ja teknologisiin taitoihin, joita Big Datan hyödyntäminen vaatii.

Vaikutuksella Big Dataa kuvaavana teemana De Mauro ym. (2015) tarkoittavat sitä, että Big Datan käytöllä ja hallinnalla on monenlaisia vaikutuksia yhteiskunnassamme. Heidän mukaansa voidaan osoittaa, että Big Dataa hyödyntävät ratkaisut ovat sopeutumiskykyisiä erilaisten vaatimusten ja alojen suhteen. Ongelmia, joita yhteiskunnan eri osa-alueilla esiintyy voi olla mahdollista ratkaista hyödyntämällä samoja datatyyppejä ja tekniikkoja. Yhtenä esimerkkinä tällaisesta De Mauro ym. (2015) pitävät Googlen hakupalvelun pohjalta tehtävää analyysia, jota voidaan hyödyntää esimerkiksi lääketieteessä ja taloustieteessä. De Mauron ym. (2015) luokituksen pohjalta esiintyviä teemoja ja niihin laajemmin liittyviä aihepiirejä esitellään kuviossa 1 (KUVIO 1)



Kuvio 1. Big Dataa kuvaavat teemat ja niiden aihepiirit (mukaillen, De Mauro ym., 2015, s. 5)

De Mauro ym. (2015) pyrkivät löytämään Big Datalle sopivan määritelmän. Heidän mukaansa aiemmat tulkinnat Big Datasta voidaan jakaa neljään ryhmään. Nämä ryhmät ovat 1) Dataan liittyvät attribuutit 2) teknologiset tarpeet 3) kynnys ja 4) sosiaalinen vaikutus. Dataan liittyvillä attribuuteilla tarkoitetaan esimerkiksi aiemmin kuvattua kolmen v:n mallia (määrä, nopeus ja moninaisuus). Teknologisilla tarpeilla taas tarkoitetaan niitä tarpeita, joita suuri datamäärä aiheuttaa. De Mauron ym. (2015) mukaan Microsoft kuvailee Big Dataa prosessiksi, joka vaatii merkittävää laskentatehoa, jota käytetään erittäin massiivisiin ja usein myös kompleksisiin datasetteihin. Kynnyksellä De Mauro ym. (2015) tarkoittavat sitä kynnystä, milloin datan käsittelystä tulee mahdotonta tavallisia menetelmiä hyödyntäen. Sosiaalisella vaikutuksella De Mauro ym. (2015) tarkoittavat Big Datan vaikutusta yhteiskuntaan. Boyd ja Crawford (2012) kuvaavat Big Datan olevan ”kulttuurinen, teknologinen ja tieteellinen ilmiö” (s. 663).

Boydin ja Crawfordin (2012) mukaan Big Dataa voidaan määrittää edellä kuvatulla tavalla. Heidän mukaansa tällöin Big Data ilmiönä on vuorovaikutuksessa 1) teknologian 2) analyysin ja 3) mytologian kanssa. Heidän mukaansa teknologialla tarkoitetaan maksimaalista laskentatehoa ja algoritmista tarkkuutta datan keräämisessä, analysoinnissa, yhdistelemisessä ja vertailussa. Analyysilla Boyd ja Crawford (2012) tarkoittavat suurten datamäärien tarkastelua taloudellisten, sosiaalisten, teknisten ja laillisten kaavojen löytämiseksi. Mytologialla Boyd ja Crawford (2012) tarkoittavat uskoa siihen, että Big Data itsessään tarjoaa ”korkeamman älykkyyden muotoa” ja tietoa, jota on mahdollista muuttaa oivalluksiksi, joiden tunnistaminen ja muuntaminen on ollut aiemmin ollut

mahdotonta. Boydin ja Crawfordin (2012) mukaan Big Data voidaan nähdä tehokkaana työkaluna yhteiskunnallisten ongelmien tunnistamiseen ja niihin puuttumiseen. Toisaalta heidän mukaansa Big Data on mahdollista nähdä myös eräänlaisena isoveljenä, joka rapauttaa yksityisyyden, ihmisoikeudet ja lisää hallinnon valtaa kansalaisiin nähden.

2.2 Näkemyksiä Big Datan määrittelystä

Big Datan määrittely on tunnetusti haastavaa. Kuten aiemmin mainittiin, yleisimmin tiedeyhteisössä Big Dataa on kuvattu Big Datan ominaisuuksien kautta. On kuitenkin selvää, että kyse on myös eräänlaisesta ilmiöstä, kuten Boyd ja Crawford (2012) osoittavat. Big Datan määrittelyn voidaan katsoa olevan kiinni kontekstista. De Mauro ym. (2015) ovat omassa tutkimuksessaan pyrkineet luomaan yleishyödyllisen määritelmän Big Datalle, joka ottaa huomioon niin datan ominaispiirteet, teknologiset vaateet kuin myös arvon, jota Big Datalla voidaan saavuttaa. Heidän mukaansa (s. 6) Big Dataa voisi kuvata seuraavasti: ”Big Data on tietovarallisuus, jolle on ominaista suuri määrä, nopeus ja moninaisuus, joka vaatii erityisiä teknologisia ja analyyttisia keinoja sen muuttamiseksi arvoksi”.

Big Dataa määriteltäessä on olemassa erilaisia näkökulmia. Kuten De Mauro ym. (2015) osoittavat, Big Dataa voidaan kuvailla esimerkiksi vaikutuksen, informaation, teknologian ja keinojen kautta. On kuitenkin olemassa myös muita näkökulmia ja ajatuksia Big Dataan liittyen. Manyika ym. (2011) mukaan Big Dataa voisi kuvailla dataksi, joka määrältään on niin suurta, ettei nykyteknologian kyky varastoida, hallita tai prosessoida riitä tuottamaan tarvittavaa tulosta taloudellisesti. Toinen teknologianäkökulmainen näkemys Big Datasta on Gantzin ja Reinselin (2011), jotka kuvailevat Big Datan olevan uuden sukupolven teknologioita ja arkkitehtuureita, jotka on suunniteltu erottelemaan arvoa moninaisesta suuren määrän dataseteistä tehokkaasti tarjoten korkean nopeuden tiedonkaappaamista, löytämistä ja analysointia.

Hashem ym. (2015) ovat omassa tutkimuksessaan koonneet Big Dataan liittyviä määritelmiä. Heidän mukaansa Big Dataa voisi määritellä olevan keskittymä keinoja ja erilaisia teknologioita, jotka ovat integroitu uudella tavalla tavoitteenaan paljastaa piilotettuja arvoja monipuolisesta, kompleksista ja suuresta datasetistä.

Gartnerin (2021) mukaan Big Datan määritelmä on seuraava ”Big Data on suuren määrän, suuren nopeuden ja suuren vaihtelevuuden omaava tietolähde tai omaisuus, joka vaatii uudenlaisia menetelmiä tiedon prosessointiin.”. Gartnerin (2012) mukaan nämä uudet menetelmät edesauttavat ja mahdollistavat parempaa päätöksentekoa, uusien oivallusten löytämistä sekä prosessien optimointia. Siinä missä Gartner (2021) vaatii Big Datan käsittelyyn uusia menetelmiä, kuvaa Loukides (2010) Big Dataa sen asettaman teknologisen haasteen kautta, sillä hänen mukaansa Big Datalla tarkoitetaan sitä, kun datan koko it-

sessään aiheuttaa suuria ongelmia tavanomaisille datan käsittelyyn tarkoitetuille teknologioille.

Dominique ym. (2016) menevät Big Datan määrittelyssä enemmän kohti Big Data analytiikkaa. Heidän mukaansa Big Data on itsessään raakaa dataa, joka voi olla joko rakenteellista tai ei-rakenteellista, joka on useimmiten yhdistelmä erilaisia formaatteja dataa valmiina käsittelyä, säilöntää ja käyttöä varten.

Akokan, Comyn-Wattiaun ja Laoufin (2017, s. 106) mukaan Big Data on ”termi, joka kuvaa suurten ja kompleksien datamäärien varastointia ja analysointia käyttäen erilaisia teknologioita, kuten NoSQL ja MapReduce”. Tässäkin on hyödyllistä havaita, että Akokan ym. (2017) näkemys eroaa muista näkemyksistä ja edustaa Dominiquen ym. (2016) tavoin Wallerin ja Fawcettin (2013) mukaan Big Datan voisi katsoa olevan erityisesti muotisana. Kuitenkin heidän mielestään Big Datan mukana tulee erilaisia mahdollisuuksia muuttaa esimerkiksi liiketoimintamallia tai analysoida päivittäistä liiketoimintaa. Wallerin ja Fawcettin näkökulman voi havaita eroavan muista näkökulmista merkittävästi, sillä heidän tarjoamansa näkökulma ei ota kantaa esimerkiksi Big Datan ominaispiirteisiin tai ominaisuuksiin, saati käytettyihin teknologioihin tai niihin liittyviin ongelmiin. Heidän näkökulmansa keskittyy enemmän liiketoimintaan ja siihen, miten Big Data itsessään voi auttaa yrityksiä tai organisaatioita menestymään entistä paremmin.

McKinsey (2011) kuvaa Big Dataa seuraavasti: ”Big Datalla tarkoitetaan datasettiä, joka kokonsa vuoksi on vaikea kaapata, varastoida, hallita ja analysoida nykypäivän teknologisia menetelmiä hyödyntäen”. On erityisen tärkeää huomata, että McKinseyn kuvaus on peräisin vuodelta 2011. Oraclen määritelmän mukaan Big Datan voidaan katsoa olevan arvotiheydeltään alhaista, tarkoittaen, että Big Datan arvo suhteessa sen määrään on erityisen alhainen.

Mayer-Schönberger ja Cukier (2013) tarjoavat Big Datalle oman määritelmänsä. Heidän mukaansa Big Datalla tarkoitetaan asioita, joita on mahdollista tehdä suuressa mittakaavassa, muttei pienessä mittakaavassa, kuten esimerkiksi havaita oivalluksia tai luoda arvoa uudella tavalla. Big Dataa hyödyntämällä voidaan Mayer-Schönbergerin ja Cukierin (2013) mukaan muuttaa markkinoita, organisaatioita, kansalaisten välisiä suhteita ja hallintoja. Heidän mukaansa Big Datan hyödyntämisen hyödyt eivät vielä pääty tähän, vaan hyödyntämisellä on heidän mukaansa rajattomat mahdollisuudet.

McAfeen ym. (2012) mukaan Big Datan avulla tavoitellaan, kuten analytiikalla yleensäkin, uuden tiedon poimintaa datasta ja sen hyödyntämistä liiketoiminnassa etulyöntiaseman saavuttamisessa. McAfeen ym. (2012) mukaan kuitenkin Big Datan voidaan katsoa erottuvan analytiikasta määrän, nopeuden ja moninaisuuden suhteen.

Opresnikin ja Taischin (2015) mukaan Big Datalla tarkoitetaan dataa, joka on 1) perinteistä yritystoiminnassa syntyvää dataa, 2) sensorien tai laitteiden toiminnasta syntyvää lokitietoa tai sensoridataa sekä 3) sosiaalista dataa, jota kertyy esimerkiksi sosiaalisen median käytöstä tai esimerkiksi evästeiden käytöstä.

Abbasin ym. (2016) määritelmä Big Datalle eroaa merkittävästi aiemmista, sillä heidän mukaansa Big Data eroaa niin sanotusta tavallisesta datasta vain neljällä tavalla – määrällä, nopeudella, moninaisuudella ja todenmukaisuudella. Heidän näkemyksensä perusteella Big Data on siis vain dataa.

Danielin (2019) mukaan Big Data on ihmisten, sovellusten ja laitteiden luomaa dataa, joka on määrältään suurta ja luonteeltaan erilaista. Danielin (2019) mukaan Big Data selittää ilmiön, joka käsittää kompleksin dynaamisen datan määrän kasvun. Hänen mukaansa Big Data käsittää määrän, nopeuden, moninaisuuden, todenmukaisuuden, varmistettavuuden sekä arvon.

Mahrtin ja Scharkowin (2013) määritelmä Big Datalle ottaa huomioon teknologian kehityksen, sillä heidän mukaansa Big Datan määritelmä on suhteellinen ja tarkoittaa yhä isompia ja isompia datamääriä ajan mittaan. Heidän määritelmänsä käsittää eri näkökulmia, kuten esimerkiksi tietoteknillisen ja sosiaalisen näkökulman. Tietoteknillisestä näkökulmasta katsoen Mahrtin ja Scharkowin (2013) määritelmä kuuluu seuraavasti: ”datasetit, jotka ovat liian suuria tavallisille tallennustavoille ja tavallisille prosessointirajapinnoille. Sosiaalisesta ja humanistisesta näkökulmasta katsottuna Mahrtin ja Scharkowin (2013) määritelmä on seuraava: ”datasettien koko aiheuttaa haasteita tutkijoille sekä sovelluksille ja laitteistoille.

Ishwarappa ja Anuradha (2015, s. 320) mukaan ”Big Data on jotain suurta ja kompleksia, jota on mahdotonta käsitellä perinteisillä järjestelmillä ja perinteisillä tallennustavoilla”. Chavanin ja Phursulen (2014) mukaan Big Data terminä on enemmänkin muotisana tai saalislause, joka käsittää suuren datamäärän sekä rakenteellista että ei-rakenteellista dataa. Datasetin suuri koko ja kompleksisuus aiheuttavat Chavanin ja Phursulen (2014) mukaan merkittäviä haasteita. Heidän mukaansa Big Dataa voidaan hyödyntää esimerkiksi yritysten päätöstenteossa, erityistesti helpottamaan ja parantamaan sitä. Osaltaan Chavan ja Phursule (2014) ovat samaa mieltä Mahrtin ja Scharkowin kanssa. Molemmat ovat sitä mieltä, että Big Data käsitteenä elää ja kehittyy jatkuvasti. Chavan ja Phursule (2014) jopa menevät pidemmälle, sillä heidän mukaansa Big Datan olevan mikä tahansa suuri määrä strukturoitua, puolistrukturoitua ja strukturoimaton dataa, jolla on potentiaalia louhittuna tietona. Tämä määritelmä voisi katsoa osaltaan romuttavan aiemmat määritelmät Big Datalle, mutta Chavan ja Phursule (2014) kuitenkin lopulta huomattavat, että Big Datasta puhuttaessa määrästä puhutaan yleensä petatavuista tai eksatavuista.

2.3 Small data

Big Datan ymmärtämistä saattaa helpottaa se, että ymmärtää miten niin sanottu tavallinen data, eli small data, tarkoittaa. Small datalle, kuten Big Datallekaan, ei ole olemassa tarkkaa määritelmää, eikä käsitettä sinänsä ole edes ollut olemassa ennen Big Dataa. Hun (2015) mukaan small data voidaan kuitenkin määritellä dataksi, jonka uskotaan tai ajatellaan ratkaisevan irrallisia kysymyksiä sellaisen datan avulla, joka on määrällisesti rajallista ja strukturoitua, yhden

instituution hallussa olevaa. Hun (2015) mukaan small datan voidaan katsoa olevan ”maailma, jonka luulemme tietävämme, maailma, jossa tieto on sellaista, jota voidaan käsitellä, analysoida ja kokea ilman tehokkaiden supertietokoneiden kyvykkyyksiä” (Hu, 2015 s. 798).

Davenportin (2014) mukaan small datana miellettävää dataa käytetään yleensä tukemaan organisaation sisäistä toimintaa, esimerkiksi liittyen hinnoitteluun tai yrityksen tarjontaan. Ero Big Datan ja small datalla onkin siis huomattava, sillä toisin kuin small dataa, Big Dataa voidaan käyttää esimerkiksi uusien liiketoimintamahdollisuuksien tunnistamisiseen.

Small dataa hyödyntävässä analytiikassa dataa puretaan, siirretään ja analysoidaan myöhempää käyttöä varten. Big Dataa hyödyntäessä tämä ei ole mahdollista, sillä dataa syntyy sellaisia määriä, että sen analysointi, siirtäminen ja purkaminen myöhempää käyttöä varten aiheuttavat suuria haasteita, sillä Big Dataa hyödyntävästä analytiikasta ei ole hyötyä tämänkaltaiseen hyödyntämiseen. Kerättävä data on määrältään, vauhdiltaan ja monimuotoisuudeltaan niin erilaista, ettei se sovellu samankaltaiseen käyttöön kuin small data (Davenport, 2014).

Ahmed, Tezel, Aziz ja Sibley (2017) vertailevat Big Dataa ja small dataa tallennusmuodon, arkkitehtuurin, datatyypin, hallinnan, datan laadun, datan käsittelyn, prosessoinnin ja tulosten analysoinnin osalta. Ahmedin ym. (2017) tekemän vertailun keskeisimmät tulokset on esitetty taulukossa 1.

TAULUKKO 1 Small Datan ja Big Datan eroavaisuuksia (mukaillen Ahmed ym., 2017)

	Small Data	Big Data
Tallennusmuoto	Keskitetty	Hajautettu
Arkkitehtuuri	sarja-arkkitehtuuri, keskitetty	Rinnakkaisarkkitehtuuri. hajautettu
Datatyyppi	Homogeeninen, staattinen, yleensä kooltaan kohtuullinen	Heterogeeninen, monilähteistä, kompleksista ja dynaamista
Datan hallinta	Tietokanta	Hallinta haastavaa, dataa yhdistellään monista lähteistä
Datan laatu	Yleensä hyvää, dokumentoitua dataa	Laatu epävarmaa
Datan prosessointi	Datalle on olennainen käyttökohde	Vaatii käsittelyä, olennaisen tiedon löytäminen haastavaa

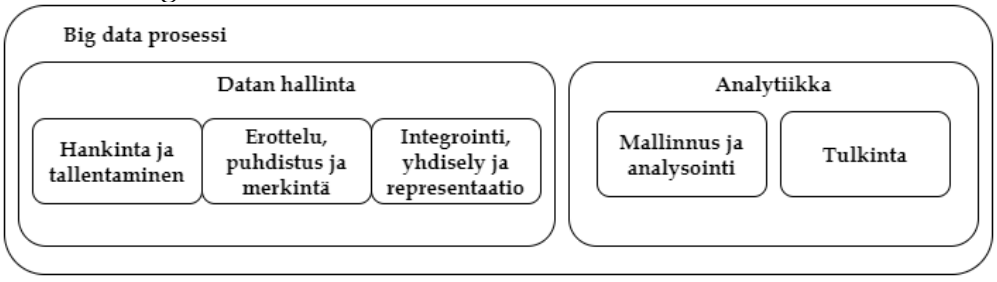
(jatkuu)

		(jatkuu)
Analyysin tulokset	Tilastollista tietoa, joka vastaa tarkkaan asetettuun kysymykseen	Ei-tilastolliset tulokset saattavat vaikuttaa merkittäviltä datan suuren määrän vuoksi

2.4 Big Data analytiikka (Big Data Analytics)

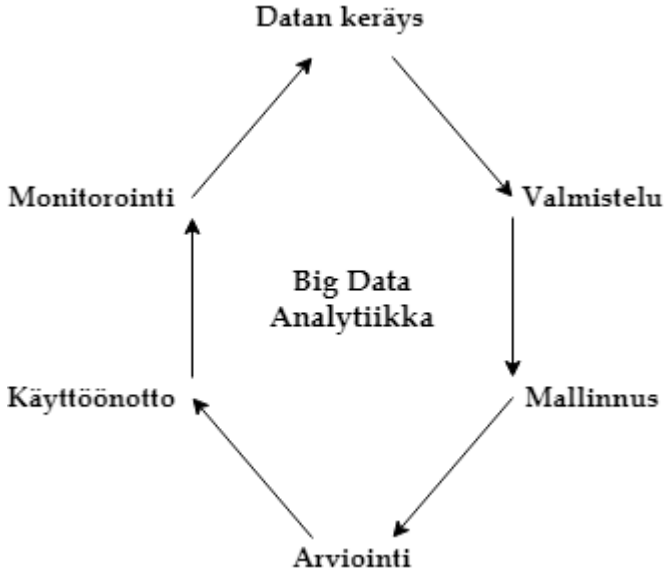
Big Data ja Big Data analytiikka menevät usein käsitteenä sekaisin. On tärkeää huomauttaa, että Big Data analytiikka on usein se, mitä eri lähteissä kuvataan Big Datana – työkalu, jonka avulla suurista datamääristä saadaan irti tietoa, jota on mahdollista hyödyntää esimerkiksi kaupallisissa tarkoituksissa. Manyikan ym. (2011) mukaan Big Data analytiikkaa voi useiden toimesta kuvata seuraavaksi edelläkävijäksi niin innovaatiossa, kilpailussa kuin myös tuotannossakin. Big Data analytiikka on saanut merkittävää huomiota niin akatemiassa kuin myös työelämässä. Rehmanin, Changin, Batoolin ja Ying Wahin (2016) mukaan Big Datan analytiikassa on kyse prosessista, jossa päämääränä on pyrkiä valjastamaan Big Dataa liiketoiminnalliseksi hyödyksi. Heidän mukaansa yritykset hyötyvät tästä prosessista, sillä se tarjoaa yrityksille uusia liiketoiminnallisia mahdollisuuksia, auttaa tunnistamaan asiakkaiden tarpeita ja helpottaa yrityksiä säilyttämään asiakassuhteita. Yritykset voivat käyttää Big Datasta louhittua analytiikkaa esimerkiksi asiakkaiden sitouttamiseen tai tukemaan tuotesuosittelua.

Gandomi ja Haider (2015) kuvaavat Big Datan analytiikka prosessina (kuvio 2), jossa suurimääräinen, nopea ja moninainen data muutetaan tarkoituksellisiksi liiketoiminnallisiksi oivalluksiksi. Heidän mukaansa prosessi koostuu viidestä vaiheesta, jotka voidaan jakaa kahdeksi alaprosessiksi – datan hallintaan ja analysointiin. Datan hallinta koostuu prosesseista ja teknologioista, joiden avulla dataa voidaan kerätä, valmistaa ja noutaa analyysia varten. Analytiikka taas viittaa erityisesti teknologioihin, joita hyödyntäen Big Dataa voidaan analysoida ja teknologioista, joiden avulla Big Datasta voidaan hankkia olennaista informaatiota. Gandomin ja Haiderin (2015) mukaan Big Datan analytiikka voidaan nähdä alaprosessina prosessille, jonka tavoitteena on ”oivallusten poiminta” Big Datasta.



Kuvio 2. Big Data analytiikan prosessi. (mukaillen, Gandomi & Haider, 2015)

Akter ja Wamba (2016) mukaan Big Datan analytiikkaa voidaan kuvata ”holistisena prosessina, joka koostuu datan keräämisestä, analysoinnista, käytöstä ja tulkinnasta” (Akter & Wamba, 2016 s. 178). Big Datan analytiikassa tavoitteena on heidän mukaansa pyrkiä valjastamaan Big Dataa liiketoiminnan avuksi siten, että se 1) edesauttaa hyödyllisten liiketoiminnallisten oivallusten saavuttamisessa 2) tuottaa liiketoiminnallista arvoja ja 3) synnyttää kilpailuetua muihin nähden. Rehman ym. (2016) ovat kuvanneet Big Datan analytiikkaa prosessina, joka koostuu kuudesta askeleesta. Nämä askeleet ovat 1) datan kerääminen 2) datan valmistelu 3) mallinnus 4) arviointi 5) käyttöönotto ja 6) monitorointi (kuvio 3).



Kuvio 3. Rehmanin ym. (2016) näkemys Big Data analytiikan prosessista (mukaillen, Rehman ym., 2016)

Datan keräämisessä on Rehmanin ym. (2016) mukaan kyse Big Datalle olennaisesta datan keräämisestä. Yritykset keräävät Rehmanin ym. (2016) mukaan dataa esimerkiksi asiakkaistaan, tuotearvosteluista, palautteista ja toimitusketjun hallinnasta. Yritysten kannalta Rehmanin ym. (2016) mukaan olennaista olisi pyrkiä luomaan strategiaa datan keräämistä varten. Datan valmistelulla Rehman ym. (2016) tarkoittavat Big Datan analytiikassa tärkeintä vaihetta, jossa olennaisessa osassa on datan laadun varmistaminen, prosessointi ja integrointi. Datan valmisteluun liittyy monenlaisia prosesseja, kuten melunvaimennus, anomalioiden havaitseminen ja poistaminen sekä mahdollisten raja-arvojen havaitseminen. Mallinuksessa tavoitteena on esimerkiksi koneoppimisen avulla pyrkiä havaitsemaan datasta mahdollisia käyttäytymiseen liittyviä kaavoja tai esimerkiksi ennustaa tulevaa. Arvioinnilla Rehman ym. (2016) tarkoittavat sitä, että mallinnuksessa mahdollisesti syntyneitä malleja tarkastellaan erilaisten keinojen avulla, jotta malli olisi varmasti sellainen, joka pystyy toimimaan myös silloin kun tuntemattoman datan määrä on maksimaalinen. Käyttöönotossa luotu malli otetaan lopulta käyttöön. Malli valjastetaan esimerkkisi yritysjärjestelmän avuksi havaitsemaan tietoa ja mahdollisia malleja ja kaavoja Big Datasta.

Monitoroinnissa tavoitteena on valvoa mallin käyttäytymistä ja toimintoja. Mahdollisen palautteen avulla aiemmin luotoa mallia voidaan hienosäätää varmistamaan, että mahdolliset tiedonsiirrot voidaan käsitellä tehokkaasti.

Big Datan analytiikka voidaan toteuttaa monin eri tavoin. Sivarajahin ym. (2017) mukaan Big Datan analytiikkaa voidaan toteuttaa kuvaavana analytiikkana, tiedusteluanalytiikkana, ennustavana analytiikkana, ohjaavana analyysina ja ennaltaehkäisevänä analytiikkana. Kuvaavalle analytiikalle on ominaista se, että sen avulla pyritään selvittämään liiketoiminnan nykyistä tilaa siten, että mahdolliset jatkokehityksen aiheet, mallit ja poikkeukset pystytään selvittämään. Kuvaavalle analytiikalle on yleistä perinteiset raportit ja tapausraportit. Tiedustelevalle analytiikalle olennaista on tutkia datan avulla sitä, onko jokin liiketoiminnallinen suuntaus mahdollinen. Kyseessä voi olla esimerkiksi jonkin tyyppinen strateginen muutos tai vaikkapa uuden tuotteen lanseeraus. Tiedustelevan analytiikan olennaisena tehtävänä on pyrkiä löytämään todisteita sille, onko esimerkiksi uuden tuotteen lanseeraus kannattavaa vai olisiko kyseinen lanseeraus jäädytettävä. Tiedusteleva analytiikka tuottaa yleensä tilastollisia analyyseja liikejohdon tueksi. Ennustava analytiikka kuvaa nimensä mukaisesti tulevaisuutta ”mitä mahdollisesti tapahtuu tulevaisuudessa?” on olennainen kysymys, johon ennustavalla analytiikalla pyritään vastaamaan. Uudet trendit ja liiketoiminnalliset mahdollisuudet ovat olennaisessa osassa ennustavassa analytiikassa. Ohjaavan analytiikan tavoitteena on optimoida liiketoiminnan prosesseja ja parantaa palvelun laatua. Samalla ohjaavalla analytiikalla tähdätään kulujen pienentämiseen. Ennaltaehkäisevällä analytiikalla pyritään liiketoiminnassa siihen, että pyritään tunnistamaan konkreettisia askelia halutun tilanteen saavuttamiseksi ja toisaalta myös varautumaan siihen, ettei haluttu tila välttämättä toteudukaan halutunlaisesti. Ennaltaehkäisevän analytiikan tavoitteena on myös pyrkiä varautumaan tuleviin ei-haluttuihin lopputuloksiin (Sivarajah ym., 2017).

Pappasin, Mikaefin, Krogstien ja Giannakosin (2018) mukaan Big Datalla tarkoitetaan ”uusia teknologioita ja arkkitehtuureita, jotka ovat suunniteltu taloudellisesti louhimaan arvoa erityisen suurista, sisällöltään vaihtelevista datamääristä mahdollistaen jatkuvan tiedon nopean kaappaamisen, uuden datan löytämisen sekä analysoinnin. On huomattava, että Pappas ym. (2017) tarjoama näkemys Big Data analytiikasta ei liiammin eroa Gantzin ja Reiselin (2011) näkemyksestä Big Dataan liittyen.

3 BIG DATAN YLEISIMMÄT OMINAISUUDET

Kuten aiemmin on saatu jo havaita, Big Datan määritteleminen on riippuvaista näkökulmasta. Suuri osa Big Dataa määrittelevistä tutkijoista turvautuu määrittelyssään Big Datan ominaispiirteisiin. Kuten todettua, Big Datan käsitteellistetään kolmen ominaisuuden, määrän, nopeuden ja moninaisuuden perusteella (Kitchin & McArdle, 2016). Myöhemmin kirjallisuudessa on tuotu lisää kuvaavia ominaisuuksia. Yleisimmin käytettyjä Big Dataa kuvaavia ominaisuuksia ovat näiden kolmen lisäksi vaihtelevuus (variability), kompleksisuus (complexity), arvo (value), todenmukaisuus (veracity) ja visuaalisuus (visualization). Näiden ominaisuuksien esiintymistä lähdekirjallisuudessa kuvaillaan taulukossa 2.

Taulukosta voidaan havaita, että vuosien edetessä perinteiset kolme ominaisuutta (3 V's) on todettu tutkijoiden mielestä riittämättömiksi. Uusimmissa akateemisissa papereissa Big Dataa kuvaavia ominaisuuksia ehdotetaan jopa 17. On siis selvää, että kattavalle tulkinnalle Big Datasta on merkittävä tarve. Toisaalta teknologiset edistysaskelet ovat varmasti myös vaikuttaneet siihen, mikä määrittää Big Dataa. Osaltaan vaikuttaa siltä, että Big Datan määrittely on lähtenyt suorastaan laukalle. Kaupalliset osapuolet ovat edesauttaneet Big Datan määrittelyn vääristymää, sillä uusia ominaisuuksia ja ominaispiirteitä on keksitty edistämään oman ratkaisun menekkiä. Osaltaan Big Datan määrittely on saanut myös humoristia piirteitä, sillä verkon keskustelupalstoilla on keksitty uusia kuvaavia ominaisuuksia Big Datan määrittelyn avuksi miltei jatkuvalla syötöllä, osin tosissaan ja osin enemmänkin humoristisessa tarkoituksessa (Shafer, 2017).

TAULUKKO 2. Big Datan ominaisuuksien ilmeneminen lähdekirjallisuudessa

Lähde	**Määrä**	**Nopeus**	**Monin.**	**Vaiht..**	**Kompl.**	**Arvo**	**Todenmuk.**	**Vis.**
Basha ym. (2019)	x	x	x	x		x	x	x
Cavanillas, Curry & Wahlser (2016)	x	x	x			x	x	
Chen ym, 2014	x	x	x			x		
Fredriksson ym. (2017)	x	x	x			x	x	
Gandomi & Haider, 2015	x	x	x	x		x	x	
Gupta & Rani (2019)	x	x	x	x	x	x	x	
Hariri, Fredericks & Bowers (2019)	x	x	x			x	x	
Ishwarappa & Anuradha (2015)	x	x	x			x	x	
Kaisler ym. (2013)	x	x	x		x	x		
Lyko, Nitzschke & Ngonga Ngomo (2016)	x	x	x			x		
Seddon & Currie (2017)	x	x	x	x		x	x	x
Ularu ym. (2012)	x	x	x				x	
Yaseen & Obaid (2020)	x	x	x	x	x		x	
Özköse, Ari & Gencer (2015)	x	x	x			x	x	

3.1 Määrä

Big Dataa määrittävistä ominaisuuksista ehkä yleisimmin käytetty ja hyväksytty on datan määrä (volume). Termissä Big Data voidaan jo havaita, että datan määrä on suurta (Katal, Wazid & Goudar, 2013). Määrällä Big Dataa määritellessä viitataan merkittävään määrään saatavissa olevaan, tallennettuun, kerättyyn ja analysoituun dataan, eikä sille ole olemassa kiinteää määritelmää. Haririn ym. (2019) mukaan dataa voitaisiin pitää Big Datana, jos datamäärä olisi eksatavun (10^{18}) ja tsettatavun (10^{21}) välimaastossa. Nykyisellään datan määrä on jo merkittävä, sillä puhutaan teratavuista ja petatavuista, mutta ennustetaan jo myös, että datan määrän jatkaessa räjähdysmäistä kasvuaan pian puhutaan jo tsettatavuista. Eatonin ym. (2012) mukaan vuonna 2011 datan määrän osalta puhuttiin noin 1,8 tsettatavusta. Samalla ennustettiin, että vuonna 2020 saatavilla olevan datan määrän oletetaan kasvavan ainakin viisinkymmenkertaiseksi aiemmasta tasosta. Joissakin arvoissa puhutaan jopa jottatavuista (10^{24}). Eatonin ym. (2012) mukaan datan määrän räjähdysmäisen kasvun takana on etenevä digitalisaatio ja uudet teknologiset innovaatiot, jotka mahdollistavat datan keräämisen uusilla tavoilla.

Katalin ym. (2013) mukaan datan määrä on olennainen ominaisuus Big Dataa määritellessä. Suurella määrällä dataa pystytään tunnistamaan esimerkiksi kuluttajakäyttäytymistä ja saavuttamaan uudenlaisia oivalluksia data analytiikan avulla. Yhä useammat organisaatiot ovat rohkaistuneet arvioimaan Big Datan avulla omaa liiketoimintaansa. Ishwarappa ja Amaradha (2015) yhtyvät Katalin ym. (2013) mielipiteeseen siitä, että datan määrä on olennainen ominaisuus – heidän mukaansa juuri datan määrä on ominaisuuksista ensimmäinen, joka tulee ihmisten mieleen, kun he ajattelevat Big Dataa.

Yleisesti määrästä puhuttaessa viitataan siihen, että datan määrä on niin merkittävä, ettei sitä pystytä käsittelemään perinteisiä menetelmiä hyödyntäen (Manyika ym., 2011). On kuitenkin tärkeää huomata, ettei datan merkittävä määrä ole tekijä, joka tekee datasta Big Dataa. Useampi määritelmä Big Datasta määrittää Big Datan olevan ominaisuuksiltaan vähintään nopeasti syntyvää ja monista lähteistä koostuvaa. Kaisler, Armour, Alberto Espinosa ja Moneyn (2013) mukaan datan määrällä tarkoitetaan sitä tiedon määrää, joka on organisaation tavoitettavissa. Kaislerin ym. (2013) mukaan organisaation ei välttämättä tarvitse omistaa kaikkea dataa, vaan sillä tulisi olla ainakin pääsy siihen.

3.2 Nopeus

Nopeudella tarkoitetaan Big Datan osalta uuden datan generoitumisen nopeutta. Uudet teknologiset ratkaisut, kuten Internet of Things (IoT) tuottaa merkittäviä määriä dataa, esimerkiksi sensoreista, koko ajan. Toisaalta nopeudella Big Dataa määriteltäessä tarkoitetaan myös sitä nopeutta, jolla kerättyä tai kertynyttä dataa pystytään analysoimaan ja prosessoimaan. Haririn ym. (2019) mukaan on erityisen tärkeää, että datan prosessoinnin nopeus vastaa uuden datan generoitumisen nopeutta. Jos näin ei ole, ongelmia esimerkiksi terveydenalalla voivat olla kohtalokkaita. Nykyään dataa kertyy monista eri lähteistä, esimerkiksi modernissa autossa on lähemmäs 100 erilaista sensoria generoimassa dataa. New Yorkin pörssin kerrotaan tuottavan arviolta yhden teratavun verran dataa yhden sijoituspäivän aikana ja Wal-Martin kerrotaan tuottavan dataa miljoonan transaktion tuntivauhtia (Sivarajah ym., 2017). Datalla itsellään on ilmeisen lyhyt elinaika, joka asettaa haasteensa sen analysoinnille ja hyödyntämiselle. Sun, Strang & Rongping (2018) lisäävät, että datan nopeutta voidaan pitää jopa tärkeämpänä ominaisuutena kuin volyymia. Tämä johtuu McAfeen ym. (2012) mukaan siitä, että reaaliaikainen tai lähes reaaliaikainen data mahdollistaa organisaatioiden ja yritysten toimia entistä ketterämmin verrattuna kilpailijoihinsa. Ishwarappan ja Amaradhan (2015) mukaan nopeus viittaa sekä uuden datan syntynopeuteen että siihen, kuinka nopeasti syntynyt data saadaan prosessoitua, tallennettua ja analysoitua.

3.3 Moninaisuus

Datan muoto vaihtelee merkittävästi. Big Datan osalta ominaisuus kuvaa yhtä lailla datan ilmenemisen rikkautta (Kaisler ym., 2013). Yhä enenevissä määrin erilaista tietoa kerätään esimerkiksi sensoreista tai sosiaalisen median palveluista. Dataa voidaan kerätä paikkatiedoista tai esimerkiksi matkapuhelimista. Digitalisaation myötä yhä useammat liiketoiminnan alat digitalisoituvat ja uusia informaation lähteitä syntyy sitä mukaa.

Aiemmin datasta puhuttaessa on yleensä keskitytty esimerkiksi tietokantoihin tai laskentataulukoihin. On selvää, että nykyään datan kirjo on entistä laajempi. Nykyään datasta puhuttaessa voidaan tarkoittaa esimerkiksi kuvia, sähköposteja, videoita, audiota, tekstiviestejä ja PDF-tiedostoja. Data on siis monimuotoista. Yleisesti dataa voidaan jakaa kolmeen eri alalajiin sen rakenteen perusteella. On olemassa dataa, jolla on helposti tunnistettava rakenne, dataa, jolla ei ole tunnistettavaa rakennetta, mutta datassa olevat elementit, kuten tagit voivat edesauttaa datan järjestelemisessä, ja dataa, jolla ei ole minkäänlaista tunnistettavaa rakennetta. Yleisimmin näitä alaluokkia kutsutaan strukturoiduksi dataksi, semi-strukturoiduksi dataksi ja strukturoimattomaksi dataksi (Kaisler ym., 2013; Hariri ym., 2019).

3.3.1 Strukturoitu data

Strukturoidulla datalla tarkoitetaan dataa, jolla on olemassa tietty rakenne. Tämä edesauttaa sitä, että dataa tai sen osasia voidaan kohdistaa, organisoida ja saavuttaa erilaisissa yhdistelmissä siten, että datasta on enemmän hyötyä organisaatiolle. Strukturoitua dataa on helppo jäsentää. (Hariri ym., 2019). Daven ja Kamalin (2017) mukaan tällainen strukturoitu data on löydettävissä esimerkiksi tietokannoissa, organisaatioille tarkoitetuissa ratkaisuissa kuten relaatiotietokannoissa ja tietovarastoissa.

3.3.2 Semi-strukturoitu data

Semi-strukturoidulla datalla tarkoitetaan dataa, jolla ei ole tarkoituksellista tai tuottavaa rakennetta kiinnitettynä. Semi-strukturoitu data ei ole yleensä varastoitu minkäänlaiseen arkistoon kuten tietokantaan. Olennaista tällaiselle datalle on kuitenkin se, että siitä on havaittavissa tietoa, kuten metadataa, joka tekee datan prosessoinnista helpompaa. Semi-strukturoitu data sisältää tageja, joiden avulla eri elementtien erottelu toisistaan on helpompaa. Esimerkkejä semi-strukturoidusta datasta ovat esimerkiksi XML-dokumentit, JSON-dokumentit ja BibTex -tiedostot (Hariri ym., 2019).

3.3.3 Strukturoimaton data

Strukturoimattomalla datalla tarkoitetaan dataa, jolla ei ole tunnistettavaa rakennetta. Tällaista dataa on esimerkiksi kuvat, sosiaalisen media data, kuten paikkatieto ja metatieto. Lisäksi sensoridata, tieteellinen data, videot, kuva-arkistot, indeksoidut internet-haut, geneettinen tieto, terveystiedot, rahaliikenteestä kertyvä data sekä web-logit ovat esimerkkejä strukturoimattomasta datasta. Toisaalta on myös niin, että joillakin strukturoimattomaksi dataksi tunnistetulla datalla on olemassa tietynlainen rakenne. Arviolta noin 80 prosenttia generoidusta datasta on strukturoimatonta. Datan moninaisuus asettaa haasteita datan analysoinnille ja keräämiselle, sillä erilaiset analysointiin tarkoitetut järjestelmät eivät välttämättä ole yhteensopivia useiden datatyyppien suhteen yhtä aikaa (Hariri ym., 2019).

3.4 Vaihtelevuus

Vaihtelevuudella Big Dataa kuvaillessa tarkoitetaan pääasiassa kahta asiaa. Ensinnäkin vaihtelevuudella tarkoitetaan datan tarkoituksen muuttumista. Tällä tarkoitetaan Sivarajahin ym. (2017) mukaan sitä, että datan merkitys muuttuu jatkuvasti. Kontekstin merkitys vaikuttaa siihen, mikä merkitys datalla on. Voi esimerkiksi olla, että yhdellä sanalla yhdessä melkein samassa twiitissa voi olla täysin eri merkitys. Tätä esiintyy erityisesti datassa, joka vaatii kielellistä pro-

sessointia. Voi esimerkiksi olla, että jollakin sanalla on eri merkitys tietyissä yhteyksissä. Puhutaan siis myös kontekstista. Voi myös olla, että ajan kuluessa sanan merkitys muuttuu ja vanhoja merkityksiä tätä mukaa tulee unohtaa. Erityisesti Big Dataan pohjautuvassa analytiikassa konteksti on merkittävässä roolissa. Tällöin aiemmin kuvattu esimerkkitilanne twiitista voi toteutua ja tällöin käytettävän algoritmin kyvykkyys nousee suureen rooliin, sillä sen tulee tunnistaa sana ja sen konteksti oikein. Jos näin ei tapahdu, twiitin sanoma tai tarkoitus voi muuttua perusteellisesti, tuloksena voi olla vioittunutta tai meluisaa dataa, jonka hyödyntämisestä tulee mahdotonta (Sivarajah, 2017).

Vaihtelevuutta käytetään myös kuvaamaan Big Datan virtauksen nopeutta ja sen muutosta. Välillä kerättävissä olevan datan nopeus on korkeaa, jolloin dataa kertyy merkittävästi ja toisinaan nopeus laskee tasolle, jossa Big Datan hyödyntäminen voi olla haastavaa. Näissä tilanteissa, erityisesti nopeuden ollessa korkealla, korostuu yrityksen kyvykkyys hyödyntää ja käsitellä kerättävää dataa oikealla tavalla. (Gandomi & Haider, 2015).

3.5 Kompleksisuus

Dataa kertyy monenlaisista lähteistä, joka aiheuttaa haasteita Big Datan hyödyntämiseen. Eri järjestelmät vaativat tietynlaista dataa, jolloin syntyy tarvetta muuntaa, yhdistellä ja sovittaa eri datalähteistä kerättyä dataa. Big Datan osalta on tarpeellista yhdistää ja korreloida suhteita, hierarkioita ja useita datayhteyksiä. Erilaisten lähteiden yhdistely ja korrelointi voi osoittautua haastavaksi tehtäväksi, jolloin analysointi voi helposti lähteä hakoteille (Yaseen & Obaid, 2020). Kaisler ym. (2013) kertovat kompleksisuuden tarkoittavan mittaria, jolla mitataan datan yhteenliitettävyyttä ja keskinäistä riippuvuutta, johon jopa yksittäisen datapisteen muutos voi vaikuttaa merkittävästi.

3.6 Arvo

Arvoa pidetään ehkä tärkeimpänä Big Dataa kuvaavana ominaisuutena (Dave & Kamal, 2017). Se saattaa hyvinkin olla syy, miksi Big Dataa kerätään ja analysoidaan (Emani, Cullot & Nicolle, 2014). Se myös eroaa yleisimmin Big Datan kuvaamiseen käytetyistä ominaisuuksista, eli volyymista, nopeudesta ja moninaisuudesta siinä, ettei arvo sinänsä kuvaa Big Datan tuomia teknologisia haasteita (Hariri ym., 2019). Arvoa käytetään kuvaamaan Big Datan hyödyntämisestä koituvaa hyötyä ja käytännöllisyyttä esimerkiksi Big Data pohjaisessa päätöksenteossa sekä yleisesti mittarina datan käytettävyydelle (Kaisler ym., 2013). Monet suuret teknologiayritykset hyödyntävät Haririn ym. (2019) mukaan Big Datan aikaansaamaa arvoa omissa palveluissaan. Esimerkiksi Googlen kerrotaan hyödyntävän puhelimista kerättävää paikkatietoa oman Google Maps -palvelun parantamiseen. Eräät muut yritykset luovat arvoa Big Datan avulla

esimerkiksi mainostamisessa tai suosituksissa. Toisaalta arvo Big Datan suhteen voi myös muodostua muillakin tavoin. Esimerkiksi yritykset voivat myydä keräämäänsä dataa eteenpäin ja saada täten kerätystä datasta arvoa (Hariri ym., 2019). Erityisesti Big Data analytiikan parissa arvoa pidetään yhtenä avainominaisuutena, joka määrittää niin Big Dataa kuin myös Big Data analytiikkaa (Manyika ym. 2011).

3.7 Todenmukaisuus

Big Dataa kertyy merkittäviä määriä, joista osa voi olla vioittunutta tai esimerkiksi epätarkkaa. On kuitenkin erityisen tärkeää, että kerättävä tieto on olennaista ja laadukasta (Hariri ym., 2019). Haririn ym. (2019) mukaan IBM arvio, että huono datan laatu maksaa Yhdysvaltain taloudelle vuosittain ainakin 3,1 miljardia Yhdysvaltain dollaria. Lukoianovan ja Rubinin (2014) mukaan todenmukaisuus voidaan jakaa kolmeen alaluokkaan, jotka ovat 1) objektiivisuus 2) todenmukaisuus ja 3) uskottavuus.

Big Dataa voidaan laadukkuuden osalta luokitella kolmeen alaluokkaan, jotka ovat hyvä, huono ja määrittelemätön. Haririn ym. (2019) mukaan laadultaan hyvä data on sellaista, jonka todenmukaisuus voidaan varmistaa. Huonon laadukkuuden osalta data on tällöin sellaista, jonka luotettavuudesta ei ole takeita. Tällöin data voi olla meluisaa eli vioittunutta tai vääristynyttä. Määrittelemätön data taas on sellaista dataa, jonka luotettavuutta ei ole ainakaan vielä varmistettu.

Voidaankin todeta, että todenmukaisuus on yksi tärkeimmistä Big Dataa kuvaavista ominaisuuksista, ainakin organisaation näkökulmasta. On erityisen tärkeää pystyä varmistamaan, että kerätty data ensinnäkin on halutunlaista ja toisaalta myös todenmukaista. Jos todenmukaisuutta ei varmisteta, seuraukset korkeallakin tasolla voivat olla merkittäviä (Hariri ym., 2019).

3.8 Volatiliteetti

Big Datan volatiliteetillä tarkoitetaan sitä, kuinka kauan data on pätevää. Khanin ym. (2018) mukaan ymmärtääkseen volatiliteetin, tulee ymmärtää Big Datan määrä, moninaisuus ja nopeus. Pätevyyden lisäksi volatiliteetillä viitataan Khanin ym. (2019) mukaan datan elinkaareen – kauanko data on pätevää ja kauanko sitä on aiheellista säilöä. Datan ollessa pätevää sitä on edukasta säilöä. Jos data menettää pätevyytensä, ei sillä tällöin ole arvoa eikä sitä tule täten säilöä. Volatiliteetin kannalta olennaisessa osassa on tunnistaa, milloin datasta tulee tarpeetonta. Varsinkin nykyaikana datan määrän ollessa merkittävä, on tärkeää pyrkiä siihen, että pystyttäisiin tunnistamaan liiketoiminnan kannalta olennaista dataa. Yhä useammin päätöksen teko pohjautuu tosiaikaiseen dataan

ja täten onkin tärkeää pystyä tunnistamaan, milloin kerätty data on menettänyt merkityksensä. (Nasser & Tariq, 2015).

3.9 Visuaalisuus

Visuaalisuudella tarkoitetaan kerätyn ja jalostetun datan esittämistä muodossa, joka on helppo ymmärtää. Yksi tärkeimmistä tehtävistä, joita Big Datan prosessoimiseen tarkoitetulla järjestelmällä on, on muodostaa suuria määriä monipuolista dataa muotoon, joka on helppoa havainnollistaa ja jatkohyödyntää. Yksi esimerkki tällaisesta muuntamisesta on Daven ja Kamalin (2017) mukaan Big Datan muuntaminen esimerkiksi graafiseen muotoon. Yleisimmin Big Datan esittämisessä käytetään taulukoita, histogrammeja, vuokaavioita, aikajanoja sekä Venn -diagrammeja. Toisaalta on tärkeää tunnistaa se tosiasia, että dataa kertyy sellaisia määriä ja sellaisella nopeudella, että nykyaikaiset menetelmät esittämiseen ovat riittämättömiä, markkinoilla on kuitenkin olemassa kaupallisia ratkaisuja tällaisen datamäärän hyödyntämiseen.

Visualisoinnilla voidaan sanoa olevan suuri merkitys päätöksenteossa. Kuten aiemmin on jo tuotu ilmi, ettei yritysjohtaja yleensä luota päätöksenteossaan olevaan dataan. Hyvin visualisoitu ja esitetty informaatio auttaa tekemään parempia päätöksiä, sillä dataan pohjautuvat päätökset ovat todistetusti yleensä parempia kuin intuition pohjalta tehdyt päätökset. (McAfee ym., 2012).

3.10 Muita kuvaavia ominaisuuksia

Big Datan kuvaamisessa on käytetty myös useita muita osmanisuuksia. Khan ym. (2018) listaavat kaiken kaikkiaan 10 ominaisuutta Big Datalle. Aiemmin esiteltyjen lisäksi Khanin ym. (2018) mukaan Big Dataa kuvaavia ominaisuuksia ovat viskositeetti elinkelpoisuus ja validius. Viskositeetin Khan ym. (2018) näkevät samankaltaisena ominaisuutena kuin aiemmin esitelty kompleksisuus. Elinkelpoisuus viittaa siihen, että Khanin ym. (2018) mielestä Big Datalla tulisi olla kyvykkyys elää ja toimia ikuisesti. Validiudella Khan ym. (2018) tarkoittavat sitä, että datan tulisi olla oikeanlaista. Tämän voisi katsoa viittaavan todenmukaisuuteen, mutta Khanin ym. (2018) mielestä näitä kahta ominaisuutta erottaa eri konseptit ja teoriat.

Opresnik ja Taisch (2015) kuvaavat Big Dataa viiden v-kirjaimen avulla. Neljä näistä, määrä, vaihtelevuus, nopeus ja arvo ovat tunnistettuja myös monien muiden tutkijoiden toimesta. Näiden neljän lisäksi he kuvaavat Big Dataa viidennellä v-kirjaimella, verification – todentaminen, vahvistus. Heidän mukaansa tätä viimeistä v-kirjainta Big Datan kuvailemisessa on käyttänyt erityisesti Beulke (2011). Opresnikin ja Taischin (2011) mukaan Beulke tarkoittaa vahvistuksella sitä, että suurten datamäärien joukossa on väistämättä dataa, joka on laadultaan tai tarkoitusperältään huonoa. Beulken (2011) mukaan vah-

vistukseen liittyy myös tietoturva, sillä hänen mukaansa on toisaalta tärkeää varmistaa eri osapuolien oikeanlaiset oikeudet.
Guptan ja Ranin (2019) mukaan yleisimpien ominaisuuksien ja ominaispiirteiden lisäksi Big Data kuvaamaan voidaan käyttää validiivisuutta, jolla Gupta ja Ranin (2019) tarkoittavat datan oikeanlaisuutta juuri tarvittavaan käyttöön. Oikeanlaisuuden varmistaminen on Guptan ja Raninin (2019) mukaan yksi työläimmistä ja kriittisimmistä tehtävistä ennen kuin Big Datasta voidaan heidän mielestään saada arvoa. Yleisimmät esitellyt ominaisuudet ja niiden osalta yhteenveto on esitelty taulukossa 3.

TAULUKKO 3. Yhteenveto omaisuuksien kuvauksesta.

Ominaisuus	**Yhteenveto**
Määrä	Data on määrältään niin suurta, ettei sitä voida käsitellä tavallisia menetelmiä hyödyntäen. Määrä itsessään on suuri, puhutaan tsettatavuista.
Nopeus	Dataa syntyy merkittävällä nopeudella, joka aiheuttaa haasteita järjestelmille. Dataa tulisi analysoida ja tallentaa samankaltaisella nopeudella kuin uutta dataa syntyy.
Moninaisuus	Data ei ole samankaltaista, vaan muodoltaan ja päätteiltään vaihtelevaa. Data voi olla rakenteellista, eli esimerkiksi tietokantoja, tai ei-rakenteellista, kuten videoita tai kuvia.
Vaihtelevuus	Uutta dataa syntyy tahdilla, joka ei ole vakio, jolloin Big Datan hyödyntämine voi olla haastavaa sekä silloin, kun dataa virtaa merkittävästi, kuin myös silloin kun dataa virtaa erityisen vähän. Vaihtelevuus viittaa myös siihen, että datan tarkoitusperä vaihtelee.
Kompleksisuus	Dataa kertyy erilaisista lähteistä. Data voi olla erilaista, jolloin sen hyödyntäminen on haastavaa perinteisillä menetelmillä. Erilaisten datalähteiden yhdistely ja suhteiden löytäminen voi olla haastavaa.
Arvo	Liiketoiminnan kannalta tärkeä ominaisuus, eroaa muista koska ei välttämättä aiheita teknologista haastetta. Arvon määritelmä voi vaihdella riippuen kontekstista.
Todenmukaisuus	Osa datasta voi olla vioittunutta tai vääristynyttä. Datan tulisi olla laadukasta ja olennaista. Big Datan hyödyntämisen kannalta yksi olennaisimmista ominaisuuksista.
Volatiliteetti	Datan pätevyyden määritelmä, mittaa sitä, kuinka kauan kerätty data on pätevää. Tärkeää tunnistaa, milloin datasta tulee tarpeetonta. (jatkuu)
	(jatkuu)

Visuaalisuus	Kerätyn datan esittämistä muodossa (esimerkiksi graaffinen), jossa se on helppo ymmärtää ja jatkohyödyntää. Merkittävä vaikutus erityisesti päätöksenteossa.

4 METODOLOGIA

Tässä luvussa keskitytään tutkimuksen toteutuksen kuvailemiseen. Aluksi esitellään valittu tutkimusmenetelmä, eli laadullinen tutkimus. Sen jälkeen kuvaillaan tutkimuksen taustaa ja tiedonkeruumenetelmää ja kuvaillaan järjestettyä kyselyä ja sen sisältöä. Kuvaillaan myös vastaajia ja heidän taustaansa. Lopuksi esitellään valittu aineiston analysointimenetelmä ja kuvataan analysoinnin kulkua.

4.1 Tutkimusmenetelmä

Tutkimus toteutettiin laadullisena tutkimuksena. Tutkimusmateriaali kerättiin verkkokyselyn avulla hyödyntäen webropol-järjestelmää. Verkkokyselyyn tiedonkeruumenetelmänä ei ole aina ongelmaton. Verkkokyselyyn vastaaminen on helppoa ja toisaalta myös vastaamatta jättäminen on yhtä lailla vaivatonta. Kirjallisuudessa onkin tunnistettu, että liittyen verkkokyselyyn vastausprosentti jää yleensä merkittävästi alhaisemmaksi kuin tavanomaisemmilla laadullisilla tiedonkeruumenetelmillä (Nulty, 2008).

Verkkokyselyn avulla on mahdollista tavoittaa merkittäviä väkijoukkoja kyselyä varten. Juuri tämän vuoksi ja vallitsevan COVID19-pandemian vuoksi päädyttiin juuri tähän menetelmään. Tutkimuksen tavoitteisiin ja tutkimusmenetelmään myös valittu tiedonkeruumenetelmä sopi hyvin. Verkkokysely sopii hyvin esimerkiksi kyselyihin, joissa tavoitteena on esimerkiksi kerätä näkemyksiä, jotka eivät välttämättä kaipaa jatkokysymyksiä (Couper & Miller, 2008). On selvää, että esimerkiksi haastatteluiden järjestämisen vaatima työ, molemmilta osapuolilta, sekä tutkimusongelma puhuvat sen puolesta, ettei tässä tutkimuksessa ole hyödyllistä hyödyntää minkään sorttisia haastattelumenetelmiä. Yleisesti voidaan todeta, että tutkittaessa ilmiöitä ja sen selityksiä, on tutkimuksen kannalta kannattavampaa valita menetelmäksi laadullinen eli kvalitatiivinen menetelmä. Määrällinen tutkimuskin voisi soveltua tähän asetettuun tutkimusongelmaan, mutta toisaalta määrällisellä tutkimuksella päästäisiin vain yleiselle

tasolle, jolla erityyppiset ymmärtämiset ja käsittämiset jäisivät auttamatta paitsioon ja vaille tulkintaa.

4.2 Tutkimuksen tausta

Kuten on jo todettu, datan määrä on kasvanut räjähdysmäisesti viimeisten vuosien aikana. Erilaiset menetelmät kehittyvät jatkuvasti ja helpottavat massiivisten datamassojen analysointia ja keräämistä. Ei kuitenkaan ole varmaa, että ihmiset, jotka saattavat työskennellä Big Datan parissa, ymmärtävät edes, mistä Big Datassa pohjimmiltaan on kyse. Yleisesti hyväksytyn käsitteen puute ei edesauta asiaa. Yhä useammat kaupalliset yritykset pyrkivät käsitteellistämään Big Datan sellaisella tavalla, joka sopii heidän omiin kaupallisiin tarkoituksiinsa, aiheuttaen lisääntyvää hämmennystä ja väärinymmärrystä. Myös akateemisten tutkijoiden erilaistavat näkökulmat ja alati lisääntyvät määrittävät ominaisuudet aiheuttavat varmasti hämmennystä.

Tässä tutkimuksessa tarkoituksena onkin paneutua juuri niihin väärinymmärryksen ja väärinkäsittämisen piirteisiin, pyrkimyksenä toisaalta löytää yleisimpiä väärinymmärryksiä ja -käsittämiä, mutta toisaalta myös pyrkiä löytämään yhteisiä piirteitä, joiden avulla Big Datan käsitteellistäminen olisi helpompaa.

Tutkimuksen tiedonkeruumenetelmänä käytettiin siis verkkokyselyä. Verkkokysely toteutettiin kahdessa osassa, siten, että toinen osa kyselystä lähetettiin Suomessa toimiville yrityksille ja organisaatioille. Toinen osa toteutettiin Jyväskylän yliopiston informaatioteknologian tiedekunnassa opiskelijoiden keskuudessa, hyödyntäen yleistä postituslistaa.

Opiskelijoille ja yrityksille lähetetyt kyselyt poikkesivat hieman toisistaan. Opiskelijoilta kysyttiin opiskeltavaa tutkintoa (kandidaatin vai maisterin tutkinto) sekä pääainetta. Yrityksille ja organisaatioille lähetetyssä kyselyssä haluttiin vastauksia titteliin sekä alaan, jolla työskennellään. Muuten kyselylomakkeet olivat identtisiä ja seuraavanlaisia:

1) Kokemus Big Datasta?
    a. Ei lainkaan
    b. Vähän
    c. Jonkin verran
    d. Paljon
2) Miten määrittelisit Big Datan? Miten se eroaa tavallisesta datasta?
3) Onko Big Datalla mielestäsi joitain ominaisuuksia tai ominaispiirteitä? Kuvaile.

Vastauspyyntöjä lähetettiin yhteensä 167 eri yritykseen ympäri Suomen. Osaltaan pyyntöjä lähetettiin suurimpiin yrityksiin, jotka toimivat esimerkiksi kaupan, mainonnan tai informaatioteknologiaan liittyvällä alalla. Lisäksi hyödynnettiin verkosta löytyviä yritystietokantoja, joista pyrittiin löytämään yrityksiä ja organisaatioita, jotka hyödyntävät Big Dataa ja myös tarjoavat siihen liittyviä palveluita osana liiketoimintaansa.

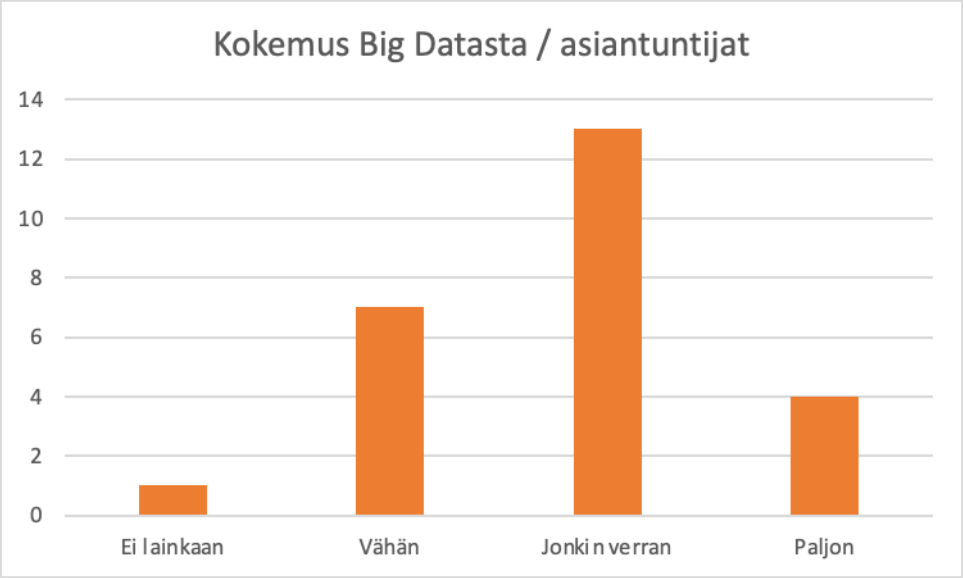
Kyselyyn vastasi lopulta yhteensä 82 henkilöä, joista 57 oli opiskelijoita sekä 25 tutkimusaiheen parissa työskenteleviä. Vastausaineistoa läpikäydessä jouduttiin suodattamaan kolme vastausta pois lopullisista vastauksista, sillä ne eivät vastanneet annettuihin kysymyksiin tai yleensäkään liittyneet aihepiiriin millään tavalla. Lopulta siis hyväksyttyjä vastauksia saatiin yhteensä 79 kappaletta.

Big Datan parissa työskennelleiden alat ja toimenkuvat vaihtelivat merkittävästi. Vastauksia saatiin seuraavilta aloilta: kryptovaluuttakauppa, lääketeollisuus, e-commerce, elintarviketeollisuus, kauppa, rahoitus, media, pankki, finanssiala, matkustaminen, IT, chatbot, ohjelmistokehitys, kuluttajatuotteet sekä IT-konsultointi. Näistä eniten vastauksia saatiin finanssialalta, yhteensä seitsemän vastausta.

Alojen esiintyvyys vastauksissa on kuvattu taulukossa. Suurin osa vastanneista koki oman kokemuksensa Big Datasta vastausvaihtoehdon ”jonkin verran ”arvoisiksi. Kaikki vastanneet kokivat, että heillä oli ainakin jonkinlainen kokemus Big Datasta. Yhteenveto kokemuksesta on esitetty kuviossa 4.

TAULUKKO 4. Asiantuntijoiden alat

Ala	Määrä
Informaatioteknologia	5
Data ja analytiikka	2
Ohjelmistot	2
Finanssiala	7
Matkustus	1
E-Commerce	2
Kuluttajatuotteet	1
Lääketeollisuus	1
Kauppa	1
Media	2
Elintarviketeollisuus	1



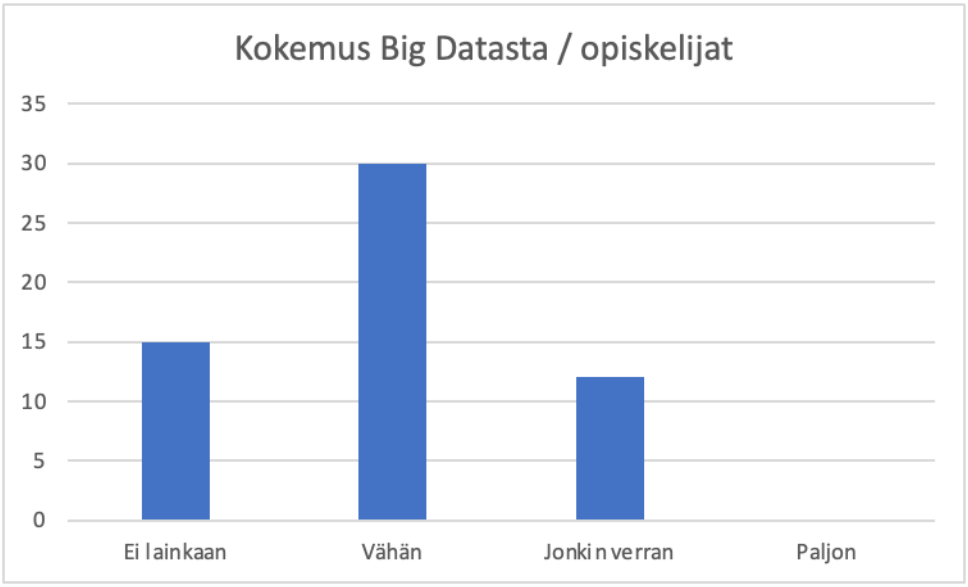
Kuvio 4. Asiantuntijoiden kokemus Big Datasta.

Vastanneiden tittelit vaihtelivat. Vastanneiden titteleitä olivat esimerkiksi service manager, data value officer, data engineer, junior data analyst, kehitysjohtaja, analytiikkajohtaja, tuotepäällikkö, toimistopäällikkö, ekonomisti sekä toimitusjohtaja. Suurin osa vastaajista työskenteli kuitenkin datan parissa tavalla tai toisella.

Opiskelijoiden keskuudessa vastauksia saatiin siis 57 kappaletta, joista jouduttiin poistamaan kolme vastausta. Suurin osa kyselyyn vastanneista oli maisterivaiheen opiskelijoita. Suuri osa vastanneista opiskeli pääaineenaan tietojärjestelmätiedettä. Opiskellun tutkinnon ja pääaineen esiintyvyyttä on kuvattu taulukossa 5. Big Dataan liittyvän kokemuksen osalta saatiin merkittävää vaihtelua, sillä vastauksia saatiin yhtä lukuun ottamatta kaikkiin vaihtoehtoihin. Kokemukseen liittyvät vastaukset on esitetty kuviossa 5.

TAULUKKO 5. Opiskelijoiden pääaineet ja opiskeltava tutkinto.

Pääaine	Määrä	Maisteri/kandidaatin tutkinto
Tietojärjestelmätiede	30	21 / 9
Kognitiotiede	30	3 / 0
Tietotekniikka	21	13 / 8
Muu	3	3 / 0



Kuvio 5. Opiskelijoiden kokemus Big Datasta.

4.3 Aineiston analysointi

Aineiston analysointimenetelmänä käytettiin laadullista sisältöanalyysia (content analysis), jonka avulla saadut vastaukset käsiteltiin. Laadullinen sisältöanalyysi valittiin aineiston analyysitavaksi siksi, että se soveltuu hyvin sellaisen aineiston käsittelyyn, jonka tavoitteena ei ole muodostaa uutta teoriaa (Hsieh & Shannon, 2005). Elon ja Kynkään (2012) mukaan laadullinen sisältöanalyysi on käytännöllinen silloin, kun tavoitteena on muodostaa tutkimuskohteesta objektiivinen ja systemaattinen kuva. He lisäävät, että sisältöanalyysi soveltuu mainiosti sellaiseen tutkimukseen, jossa tavoitteena on saada esimerkiksi teoriasta tai hypoteesista parempaa ymmärrystä. Elo ja Kyngäs (2012) ovat sitä mieltä, että laadullinen sisältöanalyysi tähtää tutkimuskeinona siihen, että sen avulla saadaan toistettavia ja päteviä päätelmiä kerätystä aineistosta niiden kontekstissa. Laadullisen sisältöanalyysin tavoitteena on heidän mukaansa tarjota kerätyn aineiston avulla uusia oivalluksia, uutta tietoa ja apua käytäntöön. Tämän vuoksi tässä tutkimuksessa hyödynnettiin juuri tätä aineiston analysointitapaa. Tutkimuksen tavoitteena ei ollut synnyttää uutta teoriaa tai niinkään tutkia olemassa olevaa teoriaa, vaan tutustua erilaisiin näkemyksiin liittyen Big Dataan ja pyrkiä tunnistamaan eri vastaajien osalta yhtäläisyyksiä, eroavaisuuksia ja myöskin lopulta käsityksiä, jotka voivat olla virheellisiä tai osaltaan puutteellisia.

Aineiston käsittelyssä hyödynnettiin koodausta, jonka avulla vastauksia pystyttiin jäsentämään ja käsittelemään helpommin. Yleisesti laadullisen sisältöanalyysin koodaus perustuu siihen, että valitaan esimerkiksi teemoja, joiden perusteella suoritetaan ensimmäinen koodaus. Tämän jälkeen suoritetaan toinen koodauskierros, jossa paneudutaan ensimmäisen kierroksen aikaansaamiin

teemoihin syvällisemmin. Tarvittaessa koodaamista voidaan jatkaa (Elo & Kyngäs, 2012).

Tässä tutkimuksessa koodaus suoritettiin seuraavalla tavalla: vastaukset koodattiin aluksi siten, että ne lajiteltiin koodin perusteella ryhmään näkökulman perusteella. Näkökulmina oli De Mauron ja Gandormin (2015) tutkimuksessa esitellyt näkökulmat Big Datan käsitteellistämiseen liittyen, eli: informaatio, teknologia, keinot ja vaikutus. Seuraavalla koodauskierroksella paneuduttiin vastauksiin syvällisemmin ja pyrittiin koodaamaan ne sen perusteella, mitä Big Datan piirteitä vastauksista löytyy (esimerkiksi määrä, nopeus, moninaisuus). Kolmannen kysymyksen kohdalla lajittelua toteutettiin siten, että vastauksista pyrittiin löytämään Big Datalle kuvaavia ominaispiirteitä, joiden perusteella vastaukset jaoteltiin omiin ryhmiinsä. Vastausten ei tarvinnut olla sanatarkasti samantyylisiä kuin kolmannessa luvussa esiteltiin.

5 TULOKSET

Tässä luvussa keskitytään analysoimaan kerättyä tutkimustietoa. Aineiston analyysissa on käytetty laadullista sisällönanalyysia. Osana sisällönanalyysia kerättyä dataa on koodattu sen mukaan, mitä näkökulmaa ne Big Datan suhteen edustavat. Näkökulmien luokittelussa on hyödynnetty toisessa luvussa esiteltyä De Mauron ym. (2015) tekemää luokitusta Big Datan näkökulmista. Tekstissä viitataan asiantuntijoihin kirjaimella A ja opiskelijoihin kirjaimella O.

5.1 Näkökulmia Big Datan määrittelystä

De Mauro ja Grimaldi (2015) määrittelivät Big Datan eri näkökulmien mukaan. Heidän mukaansa näitä näkökulmia ovat teknologia, keinot, informaatio ja vaikutus. Kuten aiemmin jo todettiin, teknologianäkökulma pitää sisällään Big Dataan liittyvät näkemykset, joiden perusteella voidaan olettaa teknologialla olevan keskiössä Big Datan määrittelyssä. Keinojen osalta keskiössä ovat keinot, joilla Big Dataa voidaan hyödyntää. Tällaisia ovat esimerkiksi koneoppiminen ja visualisointi. Informaatiokeskeinen näkökulma kuvaa sitä, millaista Big Data itsessään luultavammin on, esimerkiksi monimuotoista, järjestelemätöntä ja määrällisesti valtavaa. Vaikutus kuvaa sitä, mitä Big Datalla voidaan mahdollisesti saada aikaan ja sitä mihin sitä voidaan hyödyntää. (De Mauro ym., 2015).

5.1.1 Informaatio

Suurin osa vastanneista lähti määrittämään Big Dataa informaation kannalta. Vastaajien mukaan Big Data olisi kooltaan merkittävää tietoa, jota ei välttämättä ole järjestelty mitenkään etukäteen. Osa vastaajista uskoi, että Big Dataa määrittää vain sen suuri koko.

"Big Data eroaa "tavallisesta" datasta sen määrän suhteen" - A1

Kokoa pidettiin laajalti niin opiskelijoiden, kuin myös asiantuntijoiden, keskuudessa Big Datan tärkeimpänä ja helpoimmin tunnistettavana piirteenä. Vaikkakin datan määrästä puhuttiin laajalti, ei datan määrälle kuitenkaan tunnistettu määrettä. Erityisesti opiskelijoiden keskuudessa datan määrää kuvailtiin sanallisesti erittäin rikkain sanavalinnoin. Esimerkkejä Big Datan luokittelusta:

"Suuri määrä käsittelemätöntä tietoa" -O1

"Erittäin suuri määrä dataa" -O4

"Massiivinen könttä dataa" -O7

"Laaja datajoukko" -O10

"Big Data on järkyttävä määrä tietoa" -O13

Yksi vastaajista mielsi, että Big Dataa on paljon. Hän kuitenkin oli sitä mieltä, ettei välttämättä ole yksioikoista sanoa, mitä on paljon, sillä se riippuu hänen mukaansa kontekstista. Kuitenkin hänen mielestään Big Datan hyöty perustuu juuri datan suureen määrään, sillä sen avulla saadaan tiedolle suuri kattavuus.

"Alati kertyvää dataa, jota on määrällisesti paljon, mutta se mikä on paljon, riippuu kontekstista. Big Data voidaan nähdä sitä kautta, että se on sellaista dataa, jonka merkittävä hyöty tulee juuri sen kattavuudesta." -O19

Eräs asiantuntija koki, että Big Datan erottaa niin sanotusta tavallisesta datasta vain sen koko. Toinen taas oli sitä mieltä, että Big Dataa olisi sellainen data, jota eri järjestelmistä olisi saatu yhdistelemällä. Hänen mukaansa tällainen data olisi täysin uutta dataa.

" Mielestäni data muuttuu Big Dataksi, kun se on kokonaisuus useita erilaisia tietolähteitä / järjestelmiä mistä data syntyy kokonaisuuteen ja jonka pohjalta on tuotettu kokonaan uutta dataa." -A3

Vastaajat miettivät sitä, minkälaista dataa Big Data voisi olla. Suuri osa vastaajista oli sitä mieltä, että Big Data olisi sellaista dataa, jota on saatu monista lähteistä. Vastaajien mukaan datalla ei olisi niinkään mitään tiettyä rakennetta, vaan Big Data koostuisi erilaisista lähteistä ja erilaisista tiedostoista. Yhden vastaajan mukaan videot olisivat jo luonnostaan Big Dataa suuren määrän vuoksi, kun taas toinen vastaaja oli sitä mieltä, että Big Data koostuisi datasta, joka ei ole jokapäiväisessä tuotannollisessa käytössä. Tällä hän tarkoitti sitä, että Big Data koostuisi esimerkiksi eri järjestelmien tuottamista lokitiedostoista tai sellaisesta historiallisesta datasta, joka on saattanut aiemmin olla tuotannollisessa käytössä.

"..video on luonnostaan Big Dataa, koska sitä on paljon." -A1

" Big Data on mielestäni luonteeltaan sellaista, jota ei sellaisenaan käytetä jokapäiväisessä "tuotanto"-käytössä" -A4

" Big Data voi olla aiemmin tuotantokäytössä ollutta historiallista dataa tai alun perin Big Dataksi kategorisoitua dataa kuten järjestelmän käyttäjien tekemien toimintojen automaattisesti generoidut logitiedot." -A5

Yksi vastaajista koki, ettei Big Dataa oikeasti edes ymmärretä oikein. Hänen mukaansa Big Data käsitteenä on altis väärinkäytölle, jolla hän tarkoittaa sitä, ettei välttämättä ole ymmärrystä sille, mitä Big Data oikeasti on. Väärinkäytöksen osalta hän korostaa kirjallisuuskatsauksessakin ilmi tullutta seikkaa siitä, että kaupalliset tekijät pyrkivät muovaamaan Big Datan käsitettä muotoon, joka edesauttaa heidän kaupallisia tavoitteitaan. Vastaajan mukaan harvoilla yrityksillä itsellään olisi Big Dataa. Hänen mukaansa esimerkiksi Facebook ja Google olisivat esimerkkejä yrityksistä, jotka hallitsisivat itse Big Dataa. Suomessa hänen mielestään Big Dataa ei ole kovinkaan monilla yrityksillä. Hän mainitsee Suomen osalta, että vain media-alan suurimmilla toimijoilla saattaisi olla Big Dataa.

" Big Data käsitettä käytetään usein väärin. Googlella Facebookilla jne. on Big Dataa. Suomalaisista yrityksistä kenties media-alan suurimmilla. " -A7

Yksi opiskelijoista tulkitsi Big Datan olevan joukkio dataa ilman kontekstia, jolloin se ei vastaajan mielestä ole itsessään vielä informaatiota. Tämä näkemys oli uniikki siinä mielessä, että siinä otettiin kantaa Big Dataan informaation lähteenä. On selvää, ettei Big Datasta itsestään ole vielä mitään hyötyä, vaan sitä tulee käsitellä ja analysoida.

" Iso joukko järjestelemätöntä dataa ilman kontekstia, jolloin se ei ole vielä informaatiota." -O11

Eräs opiskelija oli sitä mieltä, ettei datan tyypille ole sinänsä mitään väliä. Hän koki, että Big Data olisi kokoelma sellaista dataa, jota saisi keinolla millä hyvänsä. Tämä kuvaa nykyajan datafikaatiota hyvin, sillä yhä useampaa asiaa voidaan mitata ja seurata teknisesti.

" Eri lähteistä kerättyä dataa kaikesta mitä saadaan vaan irti " -O12

Yksi vastaajista tiivisti omassa vastauksessaan Big Datan olennaisuuden hyvin yhteen. Hänen vastauksessaan käy ilmi, niin Big Datan luonne, siihen liittyvät ominaispiirteet sekä sen mahdollistamat tiedonkeruun menetelmät. Vastauksessa tuotiin myös ilmi, ettei Big Data sellaisenaan olisi valmista käytettäväksi, vaan vaatisi jatkokäsittelyä. Vastaajan näkemys on lähellä akateemista näkemystä.

” Data on tietoa, jota ei olla vielä käsitelty, se on vielä "raakaa" eli numeroita, kirjaimia tai muuta vastaavaa ilman että se vielä kertoo juuri mitään tutkittavasta asiasta. Big Dataa on, kun tällä datalla on volyymia, nopeutta sekä erilaisuutta. Sitä voidaan kerätä nopeasti, paljon sekä hyvin vaihtelevassa muodossa tai vaihtelevista lähteistä.” -O55

5.1.2 Teknologia

De Mauro ym. (2015) korostavat teknologian merkitystä Big Datan suhteen. Heidän mukaansa teknologia on välttämättömyys Big Datan hyödyntämisessä.

Opiskelijoiden keskuudessa teknologian merkitystä Big Datan suhteen ei laajemmin noteerattu, mutta asiantuntijat huomioivat omissa vastauksissaan myös teknologian merkityksen. Osaltaan vastauksista voidaan havaita, että Big Datalle olisi olennaista se, että sen käsittelyyn tarvitaan erityisiä teknologioita. Osa vastaajista koki, että Big Datalle olennaista olisi se, ettei sen käsittely onnistuisi yksittäistä järjestelmää tai tietokonetta hyödyntämällä. Yhden vastaajan mukaan Big Datan käsittelyn haasteellisuus yksittäisellä koneella tulee ilmi jo siinä, ettei tietokoneiden muistikapasiteetti riitä alkuunkaan käsittelemään Big Dataa.

”Big Datalla viitataan datajoukkoihin, jotka ovat liian suuria käsiteltäväksi yhdellä työasemalla, koska yksittäisen laitteen muistiin mahtuu kerrallaan vain murto-osa datakokonaisuudesta. Tämän vuoksi tarvitaan erikoistuneita klusteriratkaisuja datan käsittelyyn. Itse käsittelyyn käytettävät algoritmit tai datan luonne muilta osin eivät välttämättä sinänsä suuresti poikkea "tavallisesta" datasta” – A4

Vastauksissa kävi myös ilmi, että Big Datan käsittelyyn vaaditaan erityisiä järjestelmiä. Vastaajien mukaan tällaiset järjestelmät toisaalta mahdollistavat Big Datan käsittelyn tehokkaalla tavalla mutta myös toisaalta edesauttavat käyttäjää Big Datan hyödyntämisessä.

Teknologiakeskeinen näkökulma tuli ilmi myös opiskelijoiden vastauksissa. Yleisimmin Big Datan määrittely teknologiakeskeisestä näkökulmasta keskittyi siihen, että dataa on niin paljon, ettei sitä pystytä käsittelemään perinteisin menetelmin. Lisäksi jo aiemmin todettu datan määrän jatkuva lisääntyminen nähtiin vastaajien keskuudessa haasteena.

” datan määrä lisääntyy koko ajan ja sitä ei voi hallita esimerkiksi yhden ihmisen avulla kynän ja paperin kanssa, koska määrät ovat niin valtavia.” -O17

"Big Data on sellaisen datamäärän kerääminen ja analysointi, joka vaatii määränsä takia erityisiä toimenpiteitä"-O43

Vastauksissa tuli myös ilmi vaateita teknologisille ratkaisuille. Vastauksissa korostettiin suuren laskentatehon merkitystä sekä tallennuskapasiteetin tarvetta. Vastaus ottaa kantaa Big Datan hyödyntämiseen liittyviin haasteisiin, jotka ovat usein teknologiariippuvaisia.

" Suuri määrä kerättyä tietoa, jonka käsitteleminen saattaa vaatia paljonkin laskennallista tehoa." -O34

5.1.3 Keinot

On selvää, että datan massiivinen määrä aiheuttaa haasteita. Tarvitaankin tarpeeksi tehokkaita keinoja ja menetelmiä, jotta Big Dataa voidaan oikeasti hyödyntää. Vastauksien perusteella on selvää, että datan prosessointi vaatii oikeanlaisia työkaluja ja suunnitelmallisuutta. Yksi vastaajista oli sitä mieltä, ettei Big Dataa voisi käsitellä esimerkiksi Microsoft Excelissä.

"Datan prosessointi vaatii työkaluja, koska sitä on määrällisesti niin paljon. Datan tallentaminen ja hyödyntäminen vaatii suunnitelmallista työtä ennen kuin siitä saa kunnolla hyötyjä irti. Ei riitä, että vain "kopioi Exceliin rivejä" ja sitten sen jälkeen miettii, että mitä sille tekisi." -A17

Yksi vastaajista kokee, että Big Data oli ja meni, koska sen hyödyntämiseen ei olla onnistuttu kehittämään tarvittavia menetelmiä tarpeeksi ajoissa. Toisen vastaajan mielestä Big Datasta ei itsessään ole sinänsä hyötyä, vaan sitä tulisi käsitellä ja jalostaa. Vastauksissa korostettiin myös Big Datalle kustomoituja teknologisia ratkaisuja. Näkemys siitä, että Big Data olisi jo mennyttä, oli vastaajien keskuudessa uniikki ja erosikin merkittävästi muista Big Datan olemassaoloon liittyvistä vastauksista, sillä useampi vastaaja mielsi, että Big Dataa olisi käytössä miltei jokaisella suurella organisaatiolla.

" Useimmiten kiinnostavat löydökset ovat datasta johdettavissa monimutkaisimmissa arvoissa, eikä suoraan nähtävissä datasta tai yksinkertaisista summauksista, kuten keskiarvosta tai rivien määrästä" -A13

"Tämän vuoksi tarvitaan erikoistuneita klusteriratkaisuja datan käsittelyyn. Itse käsittelyyn käytettävät algoritmit tai datan luonne muilta osin eivät välttämättä sinänsä suuresti poikkea "tavallisesta" datasta" -A22

"Mielestäni Big Data oli ja meni. Silloin kun oli sen hype, niin sitä ei silti osattu oikein hyödyntää. Ei ollut siihen osaamista, eikä tekniset ratkaisut tukeneet Big Datan käsittelyä" -A23

5.1.4 Vaikutus

Yksi merkittävä tekijä Big Datan suhteen on sen vaikutus. Kuten todettua, Big Dataa hyödyntämällä voidaan saada paljon aikaan. De Mauron ym. (2015) mukaan Big Dataa voidaan pyrkiä kuvailemaan myös sen vaikutuksen kautta. Vaikutus voi koskettaa esimerkiksi vain yhtä organisaatiota, joka Big Dataa hyödyntämällä pystyy tehostamaan toimintaansa tai vaikutus voi koskettaa myös kokonaisia yhteiskuntia. Big Datan vaikutus voi näkyä myös esimerkiksi erilaisina sovelluksina tai myös arvona.

Vastaajien keskuudessa Big Datan määritelmää ei laajasti lähestytty sen mahdollisen vaikutuksen näkökulmasta. Yhden vastaajan mukaan Big Datalla tulisi olla jokin tarkoitusperä, jonka vuoksi dataa kerätään. Toisin sanoen, datan keräämistä tulisi ohjata jokin tarkoitusperä, esimerkiksi jonkinlainen tavoitetila. Ilman tavoitetta ei vastaajan mielestä Big Datalla ole käyttöä, vaan siitä tulee hyödytöntä.

" Big Datan keräämisellä tulee olla jokin tarkoitusperä, muuten datan kerääminen on hyödytöntä. " -O44

Toisen vastaajan mielestä taas Big Data on dataa, jota on kerätty ilman tarkoitusperää. Hänen mukaansa Big Dataa kerättäessä ei välttämättä edes vielä tiedetä, mitä datalla halutaan saada aikaan. Osa vastaajista kokee, ettei Big Datalla itsellään ole suurta vaikutusta. Heidän vastausten perusteella Big Datan katsotaan sisältävän kaiken tarvittavan, mutta sitä tulisi vastaajien mielestä muokata, järjestää ja poistaa epäolennaisuudet. Tällöin siitä olisi vasta hyötyä.

" Suurta määrää kerättyä informaatiota, jota on vain kerätty ilman mitään tiettyä tarkoitusta. Välttämättä ei ole vielä päätetty mitä datasta analysoidaan, kun sen keräys aloitetaan" -O45

"Big Data sisältää tavallaan lähes kaiken tarvittavan tiedon, mutta se ei sellaisenaan ole vielä hyödyksi. Sitä täytyisi pystyä muokkaamaan, järjestelemään ja poistamaan epäoleellinen tieto. Ominaisuutena sanoisin ehkä juuri tuon, että sisältää kaiken tiedon, niin oleellisen kuin epäoleellisen" -A18

Osalle vastaajista Big Datan ja uusien teknologisten ratkaisujen hyödyntäminen yhdessä oli keskiössä. Osa vastaajista koki, että Big Dataa hyödynnetään osana uusia älykkäitä järjestelmiä tai sellaisten luonnissa. Suuri osa vastaajista koki, että Big Data olisi osa jonkintapaista älykästä järjestelmää, jolla voidaan hyödyntää esimerkiksi koneoppimista ja tekoälyä.

" Data, jota hyödyntämällä voidaan rakentaa "älykkäitä" järjestelmiä" -A16

Big Datan hyödyntämisen suurissa organisaatioissa nousi vastaukissa pinnalle. Vastausten perusteella suurella organisaatiolla luulisi jo nykypäivänä olevan

Big Dataa. Vastauksissa tosin myös arvuuteltiin sitä, mistä näkökulmasta katsottuna näillä organisaatiolla oleva data olisi Big Dataa. Näkökulma on vastausten perusteella olennainen siksi, että nykypäiväisten menetelmien avulla organisaatioiden prosesseissa saatetaan olla tekemisissä päivittäin Big Datan kanssa ilman, että sitä erikseen havaitsisi. Nykyaikaisten menetelmien avulla voidaan vastaajien perusteella saada aikaan Big Datasta jalostettua tietoa pienissäkin muodoissa.

”Sitä hyödynnetään jo joka paikassa ja on mielestäni jo arkipäivää useimmissa suurissa organisaatioissa. Toki voi olla, eri mieltä siitä, että ovatko ne "oikeasti" Big Dataa vai ei, mutta se varmaan riippuu kuulijasta ja työvaiheesta.” -A11

5.2 Big Datan ominaispiirteet

Luvussa kolme keskityttiin tunnistamaan kirjallisuuden perusteella olennaisimpia Big Datan ominaisuuksia ja ominaispiirteitä. Näitä tunnistettiin yhteensä yhdeksän, joista yleisimpi olivat määrä, nopeus ja moninaisuus. Nämä kolme tunnetaan 3V-käsitteellä, jonka ensimmäisenä esitteli Doug Laney vuonna 2001.

Verkkokyselyssä pyydettiin vastaajia kuvaamaan Big Dataan liittyviä ominaispiirteitä ja ominaisuuksia. Suuri osa vastaajista lähtikin kuvaamaan Big Dataa luvussa kaksi esitellyillä piirteillä, mutta myös poikkeavia näkemyksiä ilmentyi. Taulukossa on esitelty luvussa kolme esitellyt ominaispiirteet, niiden esiintymismäärä vastauksissa sekä esimerkkejä siitä, miten vastaajat kuvailevat Big Dataa näiden ominaispiirteiden avulla.

TAULUKKO 6. Ominaispiirteiden esiintyvyys vastauksissa.

Ominaisuus	Määrä	Tulkinta
Määrä	68	”Valtava määrä dataa” ”Datan määrä kasvaa” ”Suuri koko” ”Datan volyymit ovat suuret”
Nopeus	22	”Sitä kertyy tyypillisesti vauhdilla” ”Dataa kerääntyy jatkuvasti kovalla nopeudella"

(jatkuu)

		(jatkuu)
Moninaisuus	24	”useista eri lähteistä koostuvaa” ”Data saapuu erilaisista paikoista” ”voi olla lähes mitä tahansa”
Vaihtelevuus	4	”datan tarkoitus vaihtelee”
Kompleksisuus	0	Ei vahvistettu.
Arvo	4	”Sisältää paljon hyödyllistä tietoa”
Todenmukaisuus	0	Ei vahvistettu.
Volatiliteetti	0	Ei vahvistettu.
Visuaalisuus	1	”Hyödyllistä tietoa jos saadaan esitettävään muotoon”

Taulukosta 6 voidaan nähdä, että vastaajien keskuudessa Big Data koetaan alkuperäisten ominaispiirteiden eli määrän, nopeuden ja moninaisuuden perusteella. Myös kirjallisuudessa nämä kolme tekijää olivat yleisimpiä. Osa ominaispiirteistä ei saanut ollenkaan vastaajilta huomiota, joten näiden ominaispiirteiden hyödyllisyys ja tarpeellisuus voidaan kyseenalaistaa, ainakin pienen otoksen perusteella.

Myös uusia ominaispiirteitä pystyttiin havaitsemaan. Asiantuntijat painottivat datan epämuotoisuutta ja Big Datan luomaa haastetta sekä järjestelemättömyyttä. Epämuotoisuudella tarkoitetaan vastausten perusteella sitä, että se on kokoelma kaikenlaista tietoa. On huomioitava, että epämuotoisuus viittaa myös moninaisuuteen, mutta epämuotoisuus kuvaa ehkä paremmin sitä, että data ei ole aina määrältään tai muodoltaan samankaltaista. Haasteen osalta vastaajat viittasivat Big Datan aiheuttamaan teknologiseen haasteeseen. Datan monimuotoisuus, alati kasvava määrä ja jatkuvalla syötöllä tapahtuva lisääntyminen aiheuttavat vastaajien mielestä suuria haasteita Big Datan hyödyntämiselle. Akateemisissa tutkimuksissa haastetta ei koettu Big Datan ominaispiirteeksi, vaan ennemminkin määrittäväksi tekijäksi.

5.3 Väärinymmärrykset

Osana verkkokyselyä vastaajia pyydettiin vertaamaan Big Dataa niin sanottuun tavalliseen dataan, ja kuvailemaan sitä, miten nämä kaksi eroavat toisistaan. Tällä asettelulla pyrittiin samaan vastaajia kuvailemaan niitä suurimpia erottavia tekijöitä. Osalle vastaajista Big Data ja tavallinen data näyttäytyivät keske-

nään samanlaisina. Yleisimmin vastaajat kokivat, että Big Data on vain tavallista dataa, jota on paljon.

Yleisin väärinymmärrys Big Datan suhteen liittyy siihen, että se olisi vain dataa, jota on paljon. Jos katsoo akateemisten tutkijoiden näkemyksiä ja vertaa niitä niin opiskelijoiden kuin myös asiantuntijoiden näkemyksiin, voidaan huomata, että useassa vastauksessa nähdään Big Datan kannalta vaillinaisia määritelmiä. Tällä tarkoitetaan sitä, että näkökulmat, kuten se, että dataa on vain paljon, ei ole vielä itsessään yksinään selittävä tekijä Big Datalle. Useat akateemiset kirjallisuudet painottavat, että Big Data koostuu kylläkin datasta, jota on paljon, mutta se on sen lisäksi vielä enemmän. On vääjäämättä pelkistetty yksinkertaistus todeta, että Big Data on vain suuri kokoelma tavallista dataa. Vastauksissa tuli myös ilmi, että jokaisella organisaatiolla olisi itsellään kerättyä Big Dataa. Tämän voidaan sanoa olevan merkittävä yleistys, sillä Big Datan hallinnoinnista on olemassa monia näkemyksiä. Osa vastaajista koki lisäksi, ettei kellään välttämättä Suomessa ole Big Dataa.

”erittäin suuri määrä dataa”-O13

”suuri määrä tietoa”-O4

”Sitä kertyy eri lähteistä valtavia määriä joka organisaatiossa” -A14

Osaltaan väärinymmärryksiä aiheuttavat alati lisääntyvät Big Datan ominaispiirteet. Tämän tutkielman kolmas luku käsittelee kirjallisuuden perusteella yleisimpiä ominaisuuksia ja ominaispiirteitä, joita Big Dataan liitetään. Aikoinaan puhuttiin yleisesti vain kolmesta ominaispiirteestä, mutta nykyään erilaisia piirteitä on pyritty tunnistamaan jo kymmeniä. Yleisesti on ollut tapana keksiä vain lisää uusia V-kirjaimella alkavia kuvaavia tekijöitä, jotka saattavat tuntua yleisesti ottaen päivänselvyyksiltä. Jo alkuperäiset kolme ominaispiirrettä aiheuttavat hankaluuksia. Haastatteluiden perusteella voidaan sanoa, että niin asiantuntijat kuin myös opiskelijatkin pohtivat sitä, täytyykö kaikkien kolmen yleisimmän ominaispiirteen täyttyä, jotta datasta tulisi Big Dataa? Jos näin on, mikä olisikaan oikeanlainen määre tarvittavalle datan nopeudelle tai määrälle? Moninaisuuden kannalta määrittäminen onkin jo hieman helpompaa.

”Ääriesimerkkinä, 1000 Teratavun suuruinen relaatiotietokanta voidaan toki ajatella melko suureksi, mutta onko se kuitenkaan 'Big Dataa' sanan perimmäisessä merkityksessä? Jos olisi kyse 1000 Teratavun suuruisesta kokoelmasta satunnaisia videotiedostoja, joita koneoppimisalgoritmilla pyrittäisiin luokittelemaan eri kategorioihin, oltaisiin ehkä jo lähempänä Big Datan perimmäistä luonnetta.” -A3

” Perinteisesti Big Dataa on käsitelty kolmen tai useamman V:n kautta (velocity, variety, volume etc.). Tämä on hyvä jaottelu viiteke-

hyksenä, mutta ei anna raja-arvoja sille mikä on Big Dataa ja mikä ei. Onko kerran sekunnissa tuleva tieto Big Dataa? Kymmenen kertaa sekunnissa? Sata kertaa sekunnissa? Lisäksi että data olisi Big Dataa, pitääkö kaikkien ominaisuuksien täyttyä vai riittääkö yksi, eli onko esim. nopeasti tuleva rakenteellinen data Big Dataa, tai toisaalta hitaasti tulevat kuvatiedostot?" A19

Datan rakenteellisuudesta ei vastausten perusteella pystytä sanomaan, liittyykö Big Datan rakenteellisuuteen mitään erityispiirrettä. Toisaalta, kuten Kaisler ym. (2013) ja Hariri ym. (2019) ovat jo todenneet, Big Dataan liittyy niin strukturoitua kuin myös strukturoimatonta dataa. Vastaajissa esiintyi kuitenkin mielipiteitä niin strukturoidun datan kuin myös strukturoimattoman datan puolesta, eikä yleistä konsensusta siitä, minkälaista dataa Big Data on, pysty välttämättä muodostamaan.

"Keskeinen ominaisuus on, että se on strukturoitua" -A24

"Big Datan ominaispiirre on se, että se on ei-strukturoitua" -A25

Yksi yleisimmistä esiintyneistä väärinkäsityksistä oli se, että käsitteet Big Data ja Big Data analytiikka sekoittuivat. Osa vastaajista koki, että Big Data käsitteenä sisältäisi myös datan analysoinnin ja käsittelyn. Osalle taas Big Data oli dataa, joka jo itsessään oli valmista käytettäväksi esimerkiksi liiketoiminnan kehittämiseen. Lisäksi osa vastaajista koki, että Big Data itsessään olisi jonkintapainen joukko ohjelmistoja, joilla käsitellään suuria määriä dataa.

" Tosi suuren, jatkuvasti lisääntyvän datan keräämistä, säilyttämistä ja analysointia" -O44

" Big Data on suuren ja jatkuvasti kasvavan tietomäärän keräämistä, säilyttämistä ja analysointia" -O50

" Joukko ohjelmistoja, joiden avulla voidaan kerätä, prosessoida ja analysoida valtavia määriä dataa" -O54

" Se riippuu vähän missä kontekstissa puhutaan Big Datasta, mutta yleistävästi ja pelkistetysti Big Data on valtavien tietomäärien keräämistä, hallintaa ja hyödyntämistä monin tavoin." -O27

Mitä Big Data sitten on, oli monelle epäselvää. Osa koki, että datasta tulee Big Dataa, kun sitä on paljon. Yhden näkökulman mukaan Big Data kattaisi kaiken tietoliikenteestä syntyvän datan. Näkökulman esittäjän mukaan datan kerääjällä ei niinkään olisi väliä, vaan suuri määrä dataa olisi Big Dataa, huolimatta siitä, kuka dataa kerää ja mihin tarkoitukseen.

" Kaikesta tietoliikenteestä syntyvä massiivinen tiedon määrä, joka pyritään järjestelemään hyödylliseen muotoon." -O33

Big Dataa yleisesti myös kyseenalaistettiin. Vastauksissa Big Datan kuvailtiin olevan ennemminkin markkinoinnin aikaansaamaa hypetystä, jota ei itsessään voi hyödyntää mihinkään. Myös muutama muu vastaaja kyseenalaisti Big Datan hyödyt ja kuvaili sitä jo menneeksi. On sinällään mielenkiintoista, että Big Data nähdään vain suosion avulla ratsastavana terminä, eikä niinkään minään todellisena asiana, josta olisi hyötyä. Näkemys siitä, että Big Data olisi jo mennyttä saattaa myös kertoa vastaajasta itsestään enemmän kuin Big Datasta.

"Big Data -käsitteenä on enemmän markkinointihöpöä kuin mitään konkreettista, vähän samoin kuin AI" -A17

Asiantuntijoiden keskuudessa väärinymmärrystä ilmeni vähemmän kuin opiskelijoiden keskuudessa. Yleisimmät väärinymmärrykset liittyivät opiskelijoiden tapaan siihen, että datan määrä olisi ainoa tekijä, joka erottaa Big Datan datasta. Kuten monet akateemiset tutkijat (esimerkiksi Mikaef ym., 2011; De Mauro ym., 2015) tuovat näkemyksissään ilmi, Big Data ei ole vain määrällinen joukko dataa, vaan se pitää myös sisällään muita ominaispiirteitä. Osaltaan asiantuntijatkin sekoittavat Big Datan analytiikkaan.

" Voidaan ajatella käsittävän määrämuotoisen datan, jos volyymit ovat suuret (esim. hallinnolliset rekisterit, kauppojen myynti jne)" -A3

"Big Data eroaa "tavallisesta" datasta sen määrän suhteen." -A1

5.4 Johtopäätökset

Empiirisen tutkimuksen perusteella voidaan sanoa, että Big Data käsitteenä on varsinkin vastanneilla asiantuntijoilla kohtuullisen hyvin tiedossa. On selvää, että sähköpostilla lähetettyyn kyselyyn vastanneet ovat valikoituneet siten, että ne edustavat yrityksissään ja organisaatioissaan sellaisia toimijoita, jotka työskentelevät datan ja joissakin tapauksissa varmasti myös Big Datan parissa. Tämän vuoksi asiantuntijoiden vastauksista on haastavaa tunnistaa toistuvia väärinkäsityksiä tai -ymmärryksiä. Osalla vastanneista asiantuntijoista tosin Big Datan määrittäminen vain määrän perusteella oli esillä. Suuri osa vastaajista kuitenkin koki, että Big Data ei olisi pelkästään dataa, jota on paljon.

Opiskelijoiden keskuudessa Big Datan määrittelyn suhteen oli selvää hajontaa. Osa vastaajista oli rehellisesti sitä mieltä, etteivät he tiedä, mitä Big Data on tai missä sitä voidaan hyödyntää. Osalla vastaajista taas oli vastauksissa havaittavissa Big Datan määrittelyyn liittyviä osasia, mutta kokonaiskuva ei välttämättä ollut hallussa.

Suuri osa luvussa 5.3. käsitellyistä yleisimmistä väärinkäsityksistä ilmeni opiskelijoilla. On selvää, ettei opiskelijalla välttämättä vielä ole tarpeeksi hyvää

käsitystä Big Datasta yleensäkään, sillä Big Dataa ei ainakaan älyttömissä määrin käsitellä opintojen aikana, Big Datan tunteminen kielii siis jonkintapaisesta alan harrastaneisuudesta.

Luvussa 3 esiteltyjä tunnetuimpia Big Datan ominaispiirteitä olivat 3V – määrä (volume), nopeus (velocity) ja moninaisuus (veracity). Nämä kolme ominaispiirrettä ilmenivät jokaisessa akateemisessa tutkimuksessa, joihin tutustuttiin kirjallisuuskatsauksen tekoa varten. Myös kyselyyn vastanneiden keskuudessa nämä kolme olivat yleisimpiä Big Dataa selittäviä omaisuuksia. Osa kolmannessa luvussa esitellyistä ominaisuuksista ei saanut vastaajien keskuudessa minkäänlaista suosiota, mutta myös uusia varteenotettavia ominaispiirteitä ilmeni, korostaen Big Dataa ilmiönä ja siihen liittyvää tietynlaista mystisyyttä. Uudet ominaispiirteet, kuten järjestelemättömyys toisaalta viittaavat läheltä yleisimpiä ominaispiirteitä. Haaste itsessään ei niinkään voi katsoa kuvaavan Big Dataa itseään, vaan sen aiheuttamaa vaatimusta järjestelmälle, jolla Big Dataa hyödynnetään.

Yllättävää ominaispiirteiden osalta oli se, ettei arvo Big Dataan liittyvänä ominaispiirteenä saanut merkittävää kannatusta. Merkittävä määrä tutkijoita toi esiin arvon omassa näkemyksessään. Big Dataa kuvaavaksi ominaisuudeksi arvo sopii toisaalta hyvin ja toisaalta myös huonosti, sillä sen voi katsoa olen niin Big Datan käytön päämäärä kuin myös Big Data analytiikan aikaansaannos.

Kolmas tutkimuskysymys käsitteli sitä, että liittyykö Big Dataan väärinkäsityksiä ja -ymmärryksiä. Vastausten perusteella voidaan sanoa, että selvästikin Big Dataan liittyy väärinymmärrystä. Yleisin väärinymmärrys on se, että Big Data olisi pelkästään suuri määrä dataa. Jos otetaan huomioon kirjallisuuskatsauksessa esitettyjä näkemyksiä Big Datasta, voidaan huomata, ettei yksikään tutkija ole sitä mieltä, että Big Dataksi riittäisi se, että dataa on paljon. On selvää, että datan määrä voi aiheuttaa haasteita sen jatkokäsittelyssä, mutta Big Data on selvästikin paljon enemmän kuin vain suuri määrä dataa. Kuitenkin osa vastaajista, eritoten asiantuntijat osasivat määrittää Big Dataa oikeanlaisesti, ottaen huomioon erilaisia näkökulmia sen suhteen ja todeten, että Big Data olisi muutakin kuin määrältään suurta dataa. Osa vastaajista lähti, osin kyselyasettelun saattelemana vertaamaan Big Dataa tavalliseen dataan. Osalle vastaajista oli selvää, että Big Data erottuu tavallisesta datasta merkittävästi, mutta osan näkemys erosta rajoittui vain datan määrään. Näkökulman rajallisuus näkyi erityisesti opiskelijoiden keskuudessa.

Toinen yleisin väärinymmärrys liittyi siihen, että Big Data käsittäisi muutakin kuin vain dataa. Vastaajien mukaan Big Data sisältäisi esimerkiksi datan analysointia, rikastamista, poistamista ja muotoilua. Kirjallisuuden perusteella Big Data ei käsitä tällaisia toimenpiteitä, vaan tällöin kyse on Big Datan analytiikasta, joka itsessään on oma käsitteensä. Miksi Big Data sitten käsitteellisellä tasolla sekoittuu analytiikkaan? Vastaajien perusteella heillä ei ole välttämättä tarkkaa käsitystä Big Datan perimmäisestä luonteesta. Osa näyttää mieltävän Big Datan suuremmaksi asiaksi kuin pelkäksi dataksi, osin siksi, että heidän mielestään Big Datan tulisi erottua selvemmin niin sanotusta tavallisesta datasta.

6 POHDINTA

Mitä Big Data on? Kysymys, johon ei ole olemassa täysin aukotonta vastausta. Onko Big Data nimensä mukaisesti suuri datajoukko, vai jotain monimutkaisempaa? On selvää, että nykypäivänä kerätään suuria määriä dataa siitä, mitä teemme, missä teemme ja millä teemme. Meitä seurataan esimerkiksi evästeiden avulla ja tietojamme kaupataan eteenpäin. Myymme yksityisyytemme, jotta pystymme käyttämään palveluita.

Ihmisistä kertyy ällistyttäviä määriä dataa päivittäin. Tuota tietoa kerätään ja jalostetaan esimerkiksi kaupallisiin tarkoituksiin. Puhutaan Big Datasta, joka käsitteenä on vähintäänkin häilyvä. Sille ei ole olemassa yksiselitteistä määritelmää, vaan toinen toistaan erilaisempia määritelmiä, usein siten, että ne hyödyttävät jotakin kaupallista palveluntarjoajaa. Riittääkö. että dataa on paljon? Vai pitääkö datalla olla muitakin ominaisuuksia, että sitä voidaan kutsua Big Dataksi? Mikä sitten on paljon, riippuu näkökulmasta. Tekniikka kehittyy jatkuvasti ja siinä samassa myös dataa kerätään entistä enemmän.

Onko Big Data itsessään oikeasti sellainen asia, josta on hyötyä, vai onko se vain suurta hypetystä saanut ilmiö, jonka suosi oli ja meni? Onko Big Datan määritelmä eri, riippuen kontekstista? Miksi ylipäätään pitää olla erikseen käsitteet tavalliselle datalle, eli ”small datalle” ja massadatalle, eli Big Datalle? Miksi ei voitaisi puhua vain yleisesti datasta, jota voidaan hyödyntää eri tavoin. On selvää, että Big Datalle ja small datalle on omat käyttökohteensa, mutta kuitenkin, yleisimmin dataa joudutaan jalostamaan ja analysoimaan ennen kuin siitä on yrityksille hyötyä. Pelkällä raakadatalla ei vielä tee mitään.

Monesti onkin niin, että käsitteenä Big Data sekoittuu analytiikan kanssa. Oletetaan, että Big Data käsitteenä kattaa kaiken tiedon keruusta sen analysointiin ja jatkohyödyntämiseen. Kirjallisuuden perusteella näin ei kuitenkaan voida sanoa olevan, sillä Big Datan analysoinnille on olemassa oma käsitteensä – Big Data analytiikka. Miksi näin on? Eikö Big Data käsitteenä voisi jo kuvata koko prosessia, jolloin sen ymmärtäminen ja selittäminen käsitteen tasolla olisi merkittävästi helpompaa.

Voitaisiin mennä jopa niin pitkälle, että alettaisiin puhua vain data-analytiikasta. On myös tärkeää kyseenalaistaa Big Datan merkitystä yrityksille.

Mitä siitä pitkässä juoksussa saadaan irti, varsinkaan verrattuna perinteisempään data-analytiikkaan? On selvää, että datan mieletön määrä aiheuttaa omat haasteensa nykyaikaisen datan hyödyntämiselle, puhumattakaan datan monimuotoisuudesta.

On kuitenkin aiheellista kyseenalaistaa Big Dataan liittyviä ominaisuuksia, joita tunnutaan keksivän jatkuvalla syötöllä lisää. Osa ominaisuuksista on perusidealtaan jo niin yleisluonnollisia, että niiden yhdistäminen Big Dataan on, jos ei mahdotonta, niin ainakin hankalaa. Esimerkiksi visualisointi, joka piirteenään voisi yhtä lailla soveltua mihin tahansa analytiikkaan, koetaan joidenkin mielestä Big Dataa kuvaavaksi tekijäksi.

Mikä lopputulema Big Datan analysoinnilla sitten on? Yhden asiantuntijan mielestä Big Datastakin tulee lopulta perinteistä dataa, joten jo sen perustella on tarpeellista kyseenalaistaa Big Datan tarpeellisuus ainakin käsitteellisellä tasolla.

Vaikkakin datan määrä on räjähdysmäisesti kasvanut datafikaation vanavedessä, on tärkeää kyseenalaistaa erilaisten Big Dataan liittyvien työkalujen merkitys, esimerkiksi se, että eroaako Big Datan analytiikassa käytettävät työkalut analysoinnin suhteen oikeasti merkittävästi tavallisen datan vastaavasta? Vai onko pohjimmiltaan kyse samankaltaisista metodeista.

Tavalliselle ihmiselle Big Datan hahmottaminen on loppujen lopuksi äärettömän vaikeaa. Jokaisella tuntuu olevan oma näkemyksensä Big Datasta, joka kielii hypestä ilmiön ympärillä. Yhteistä kuitenkin kaikille näkemyksille tuntuu olevan niiden yksinkertaisuus – dataa on paljon, se riittää. Miksi näkemykset sitten ovat näin yksitoikkoisia? Siksi, ettei tavallinen kansalainen koe olevansa tekemisissä Big Datan kanssa. Useammin näin myös on, sillä kenellä oikeasti on Big Dataa? Pankeilla? Sosiaalisen median yhtiöillä, kuten Facebookilla tai Twitterillä?

Onko Big Data itsessään sellainen, jota on trendikästä kertoa hyödyntävänsä, ymmärtämättä siitä mitään? Voin itsekin myöntää, ettei minulla ole ollut selvää käsitystä siitä, mitä Big Data oikeasti on. Olen esimerkiksi olettanut ennen, että pankkien lokitiedot olisivat Big Dataa. Big Datan ymmärtämisen kannalta olisi tärkeää pystyä tunnistamaan sille yhtenäinen käsite. Miksi nykypäivänä enää keskitytään datan määriin tai sen monimuotoisuuteen, nehän kuitenkin ovat osa tätä päivää. Olisiko tärkeämpää puhua esimerkiksi datan laadusta tai esimerkiksi datan keräämiseen liittyvistä ongelmista?

Yksi kyselyyn vastanneista asiantuntijoista oli itse sitä mieltä, ettei enää nykypäivänä ole mielekästä puhua erikseen Big Datasta ja tavallisesta datasta, sillä dataa alkaa hänen mielestään olemaan kaikissa yrityksissä merkittäviä määriä. Se, miten sitä dataa sitten hyödyntää, on se tekijä joka lopulta määrittää.

Big Datalle ja tavalliselle, small datalle, yhteistä on päämäärä. Tuo päämäärä on arvonluonti sille, joka sitä hyödyntää. Mielestäni olisi mielekkäämpää puhua data-analytiikasta erilaisine menetelmine kuin hajanaisesta joukosta käsitteitä siitä, kuinka paljon dataa on oikeasti paljon. On kuitenkin varmasti turvallista todeta, että datan määrällä ei lopulta ole suurtakaan merkitystä. Merkitystä on sillä, miten sitä dataa käyttää ja hyödyntää. Tällöin puhutaankin arvos-

ta, joka voi esiintyä monin tavoin, esimerkiksi kustannusten pienenemisenä, uutena liiketoimintana tai vaikkapa onnistuneena markkinointina.

Lopulta kuitenkin on selvää, ettei Big Data itsessään pelasta ketään. Sitä tulee osata hyödyntää ja käyttää oikein, on selvää, että huono analyysi on huono analyysi, vaikka dataa olisi saatavilla merkittävästi.

Aineiston perusteella vaikuttaa siltä, että ihmisillä on vaillinainen käsitys siitä, mitä Big Data on. Mistä tämä sitten johtuu? Jo itsessään termi Big Data on mielestäni hieman harhaanjohtava ja käsitteenä epäonnistunut. Se itsessään viittaa siihen, että dataa on paljon. Tämä selittää sen yleisen näkemyksen Big Datasta vain suurena datamääränä.

7 YHTEENVETO

Tässä luvussa luodaan yhteenveto tutkielmalle. Yhteenvedon lisäksi luvussa pohditaan tutkimuksen luotettavuutta ja toistettavuutta sekä tuodaan ilmi tutkimusaiheen kannalta mahdollisia jatkotutkimusaiheita.

7.1 Yhteenveto

Tässä pro gradu -tutkielmassa oli tavoitteena tutkia sitä, miten eri tavalla Big Dataa kuvaillaan. Tarkoituksena oli verkkokyselyn avulla kerätä vastauksia kysymyksiin, jotka koskivat Big Datan määrittelyä. Tutkimuskysymykseksi tälle tutkimukselle asetettiin ”Mitä tarkoitetaan Big Datalla?”, ”Mitä ominaispiirteitä liittyy Big Dataan käsitteenä?”, ”Miten ihmiset kokevat Big Datan?” ja ”liittyykö Big Dataan väärinkäsitystä ja väärinymmärrystä?”.

Tutkielma koostui seitsemästä pääluvusta, jotka olivat johdanto, Big Data käsitteenä, Big Datan yleisimmät ominaispiirteet, metodologia, tulokset, pohdinta sekä yhteenveto. Johdannossa motivoitiin aihepiiriä kuvailemalla Big Datan käsitteellistämisen haasteellisuutta, sekä sitä, että Big Data on saanut osakseen suurta suosiota. Big Datan suhteen on tehty paljon akateemista tutkimusta sekä se on saanut merkittävää kaupallista huomiota eri toimijoiden osalta. Big Datalta on kuitenkin puuttunut yleishyödyllinen, hyväksytty määritelmä, mikä on johtanut siihen, että eri instanssit ovat luoneet monia erilaisia, kirjaviakin, määritelmiä Big Datalle. Osaltaan tämä on sotkenut ja hankaloittanut Big Datan ymmärtämistä.

Toisessa luvussa keskityttiin Big Dataan käsitteenä. Toisen luvun aluksi esiteltiin tutkielman kannalta olennaisessa osassa oleva Big Datan luokittelu, jossa De Mauro ym. (2015) pyrkivät luokittelemaan Big Dataan liittyviä käsitteitä neljän päänäkökulman, informaation, teknologian, keinojen ja vaikutuksen, avulla. Heidän mukaansa Big Datan käsitteellistäminen pitäisi sisällään nämä kaikki näkökulmat, tavalla tai toisella. Myös toisenlaisia näkökulmia pystyttiin tunnistamaan. Esimerkiksi Boyd ja Crawford (2012) kuvaavat Big Dataa eri ta-

valla. Heidän mukaansa Big Data olisi eräänlainen ilmiö, joka pitäisi sisällään niin mytologiaa, tiedettä kuin myös kulttuuria.

Luvussa esiteltiin myös kirjallisuudessa esiintyneitä käsitteellisiä näkemyksiä Big Datan suhteen. Voitiin huomata, että paikoitellen käsitteissä oli yhtäläisyyksiä, kuin myös merkittäviä erojakin. Yleisimmin Big Dataa kuvailtiin Doug Laneyn vuonna 2001 lanseeraamilla ominaispiirteillä määrä, nopeus sekä moninaisuus. Nämä ominaispiirteet ovat omiaan kuvaamaan sitä haastetta, jota jatkuvasti kasvava datamäärä saa aikaan. Osa näkemyksistä pyrkikin käsitteellistämään Big Dataa juuri sen aikaansaamien haasteiden kautta.

Kolmannessa luvussa keskityttiin kuvailemaan kirjallisuudessa esiintyneitä yleisimpiä Big Dataa kuvaavia ominaisuuksia ja ominaispiirteitä, joita tunnistettiin kirjallisuudesta yhteensä yhdeksän: määrä, nopeus, moninaisuus, vaihtelevuus, kompleksisuus, arvo, todenmukaisuus, volatiliteetti ja visuaalisuus. Myös muita harvemmin esiintyneitä ominaispiirteitä, kuten viskositeetti ja todentaminen.

Neljännessä luvussa keskityttiin tutkielman empiirisen osuuden kuvaamiseen. Luvussa esiteltiin empiirisen osuuden tutkimusmenetelmä, tutkimuksen tausta sekä kerätyn aineiston analysoinnissa käytetty menetelmä, laadullinen sisältöanalyysi. Luvussa esiteltiin myös tutkimuksessa käytetyt verkkokyselylomakkeet, joista käytettiin kahta hieman toisistaan eroavaa, toinen opiskelijoita, ja toinen asiantuntijoita varten. Empiiriseen osuuteen saatiin lopulta hyvä määrä vastauksia, mutta vastaamisprosentti jäi kuitenkin varsin alhaiseksi varsinkin asiantuntijoiden keskuudessa, jossa vastausprosentiksi saatiin 15,4 %.

Viidennessä luvussa keskityttiin analysoimaan kerättyä aineistoa. Aineistoa analysoitiin jakamalla saatuja vastauksia De Mauron ym. (2015) esittelemien näkökulmien, informaation, teknologian, keinojen, ja vaikutuksen perusteella. Voitiin havaita, että suurin osa saaduista vastauksista keskittyi voimakkaasti kuvailemaan Big Dataa informaation näkökulmasta. Suuri osa opiskelijoilta saaduista vastauksista koski nimenomaan informaation määrää. Luvussa pyrittiin myös tunnistamaan vastauksista yleisimpiä ominaispiirteitä ja peilaamaan niitä luvussa kolme esiteltyihin ominaisuuksiin. Voitiin havaita, samalla tavalla kuin näkökulmienkin osalta, että suuri osa vastaajista kokivat erityisesti määrän vaikuttavimpana ominaisuutena Big Datan määrittelyssä. Lisäksi muita yleisimpiä ominaisuksia, kuten nopeus ja moninaisuus, oli hyvin edustettuna. Toisaalta osa esitellyistä ominaispiirteistä ei saanut vahvistusta.

Kerätyn aineiston perusteella pystyttiin myös tunnistaman yleisimpiä väärinymmärryksiä ja -käsityksiä Big Datan suhteen. Voidaan sanoa, että suurin osa saaduista vastuksista oli Big Datan käsitteen kannalta näkökulmiltaan liian yksinkertaisia. Suuri osa vastaajista määritti, kuten todettua, Big Datan vain määrän perusteella. Luvussa kaksi esiteltyjen käsitteiden perusteella voidaan sanoa, etteivät nämä määritelmät ole riittäviä määrittämään Big Dataa – pelkkä määrä ei riitä. Toisena yleisempänä väärinkäsityksenä oli Big Datan sekoittuminen analytiikan kanssa. Big Data itsessään ei sisällä minkäänlaista tiedon käsittelyä tai analysointia, vaan on puhtaasti dataa, joka aiheuttaa nykyaikaisille järjestelmille haasteita sen hyödyntämisessä. Analysointi ja arvon luonti keskit-

tyvät ennemminkin Big Datan analytiikkaan, johon merkittävä osa vastaajista viittasikin kuvaillessaan niin Big Dataa, sen ominaispiirteitä kuin myös hyödyntämiskohteita.

7.2 Tutkimuksen luotettavuus

Laadullisen tutkimuksen luotettavuutta ei voida arvioida samankaltaisin menetelmin kuin määrällisen tutkimuksen. Laadullisen tutkimuksen luotettavuutta voidaan arvioida esimerkiksi toistettavuuden avulla. Tässä tutkimuksessa hyödynnettiin anonyymia verkkokyselyä, jonka avulla vastaajat vastasivat annettuihin kysymyksiin. Vastauspyynnöt lähetettiin yhteensä 167 eri yritykseen, josta saatiin suhteellisen vähän vastauksia, joka tosin on yleistä hyödynnettäessä verkkokyselyä. Vastaajien anonymiteetin vuoksi tutkimusta ei voi sellaisenaan toistaa, eikä täten tutkimuksen laatua tai toistettavuutta voida aukottomasti arvioida.

Tutkimusta rajoitti osaltaan suhteellisen pieni vastaajamäärä, niin opiskelijoiden kuin myös asiantuntijoiden osalta. Vaikkakin vastaajien taustat ja alat vaihtelivat, ei vastaajiin saatu välttämättä tarvittavaa hajontaa, vaan vastaajien otanta saattoi jäädä vajaaksi. Myös opiskelijoiden määrä suhteessa asiantuntijoihin voidaan nähdä rajoittavana tekijänä, sillä he olivat yliedustettuna kyselyn vastaajien keskuudessa. Lisäksi opiskelijoiden osalta rajoitteena on se, että miltei kaikki vastaajat tulivat samasta tiedekunnasta ja suuri osa vastanneista opiskelijoista opiskeli samaa pääainetta.

Vastaajien tausta oli myös tutkimuksen kannalta ehkä liian samankaltainen, jonka vuoksi merkittävää hajontaa ei saatu aikaiseksi. Tutkimuskysymykset olivat tutkimusongelman kannalta suhteellisen hyvin onnistuneita, vaikkakin valittu tiedonkeruumenetelmä edesauttoi vastaajia vastaamaan kysymyksiin lyhyesti, jonka vuoksi vastausten analysointi oli ajoittain haastavaa. Vastaajat myös tapasivat yleistää vastauksiaan rajusti, jonka tosin voisi katsoa liittyvän osaltaan myös Big Dataan liittyvään ymmärryksen puutteeseen ja toisaalta myös siihen, ettei Big Data saa osakseen niin laajaa kiinnostusta, kuin mahdollisesti on aikojen saatossa povattu.

7.3 Jatkotutkimusaiheet

Big Dataa on tutkittu tähän mennessä jo merkittävästi. Kuten aiemmin jo todettiin, ei Big Dataan liittyviä väärinkäsityksiä ja -ymmärryksiä ole tutkittu miltei ollenkaan. Mielestäni olisi hyvä, jos tutkimuksessa panostettaisiin Big Datan käsitteen määrittelyyn.

Toinen Big Dataan liittyvä jatkotutkimusaihe on mielestäni Big Datan eettisyys. Dataa kerätään nykyaikana merkittäviä määriä paikoitellen jopa ilman

kenenkään lupaa. Mielestäni olisi syytä tutkia keinoja, jolla Big Datan eettisyyteen voitaisiin panostaa.

Tutkimusta voisi myös keskittää esimerkiksi siihen, miksi Big Data käsitteenä ylipäätänsä on olemassa, onko esimerkiksi kyseessä enemmänkin markkinoinnin aikaansaama muotisana vai oikeasti sellainen asia analytiikassa ja tietojenkäsittelyssä yleensäkin, joka aikaansaa suurta muutosta läpi teknologisen kentän.

LÄHTEET

Ahmed, V., Tezel, A., Aziz, Z. & Sibley, M. (2017). The future of Big Data in facilities management: Opportunities and challenges. *Facilities (Bradford, West Yorkshire, England), 35*(13/14), 725-745. doi:10.1108/F-06-2016-0064

Akoka, J., Comyn-Wattiau, I. & Laoufi, N. (2017). Research on Big Data – A systematic mapping study. *Computer Standards and Interfaces, 54*(Part 2), 105-115. doi:10.1016/j.csi.2017.01.004

Akter, S., & Wamba, S. F. (2016). Big Data analytics in E-commerce: a systematic review and agenda for future research. *Electronic Markets*, 26(2), 173-194.

Andreu-Perez, J., Poon, C., Merrifield, R., Wong, S., & Yang, G.-Z. (2015). Big Data for Health. *IEEE Journal of Biomedical and Health Informatics*, 19(4), 1193–1208. https://doi.org/10.1109/jbhi.2015.2450362

Basha, S. M., Rajput, D. S., Bhushan, S. B., Poluru, R. K., Patan, R., Manikandan, R., & Kumar, A. (2019). Recent Trends in Sustainable Big Data Predictive Analytics: Past Contributions and Future Roadmap. *International Journal on Emerging Technologies*, 10(2), 50-59.

Bates, D., Saria, S., Ohno-Machado, L., Shah, A., & Escobar, G. (2014). Big Data In Health Care: Using Analytics To Identify And Manage High-Risk And High-Cost Patients. *Health Affairs*, 33(7), 1123–1131. https://doi.org/10.1377/hlthaff.2014.0041

Bedeley, R. (2014). Big Data Opportunities and Challenges: The Case of Banking Industry. *SAIS 2014 Procedings 2.*

Beulke, D. (2011). Big Data Impacts Data Management: The 5 Vs of Big Data. Haettu 13.4.2021 osoitteesta https://davebeulke.com/Big-data-impacts-data-management-the-five-vs-of-Big-data/

Boyd, D., & Crawford, K. (2012). Critical questions for Big Data: Provocationsfor a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

Chavan, M. V. & Phursule, R. N. (2014). Survey paper on Big Data. *International Journal of Computer Science and Information Technologies*, Vol. 5 (6) , 2014, 7932-7939

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of Big Data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 68, 285–299. https://doi.org/10.1016/j.trc.2016.04.005

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A survey.Mobile networks and applications, 19(2), 171-209.

Couper, M. P. & Miller, P. V. (2009). Web survey methods: Introduction. *Public Opinion Quarterly, 72*(5), 831-835. doi:10.1093/poq/nfn066

Cox, M. & Elssworh, D. (1997). Managing Big Data for scientific visualization NASA.

Curry, E., María Cavanillas, J. & Wahlster, W. (2016). New horizons for a data-driven economy: A roadmap for usage and exploitation of Big Data in europe (1st ed. 2016). Cham: Springer Open. doi:10.1007/978-3-319-21569-3

Daniel, B. K. (2019). Big Data and data science: A critical review of issues for educational research. *British Journal of Educational Technology, 50*(1), 101-113. doi:10.1111/bjet.12595

Dave, M. & Kamal, J. (2017). Identifying Big Data dimensions and structure. (s. 163-168) *IEEE*. doi:10.1109/ISPCC.2017.8269669

Davenport, T. (2014). Big Data at work: Dispelling the myths, uncovering the opportunities. *Harvard Business Review Press.*

De Mauro, A., Greco, M., & Grimaldi,M. (2015). What is Big Data? A consensual definition and a review of key research topics. *In AIP conference proceedings(Vol. 1644, No. 1, pp. 97-104).*

Eaton, C., Deutch, T., Deroos, D, Lapis, G. & Zikopoulos, P. (2012). Understanding Big Data; Analytics for Enterprise Class Hadoop and Streaming Data. McGraw Hill

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Emani, C., Cullot, N. & Nicolle, C. (2015). Understandable Big Data: A survey. *Computer Science Review, 17*, 70-81. doi:10.1016/j.cosrev.2015.05.002

Favaretto, M., De Clercq, E., Schneble, C. O., & Elger, B. S. (2020). What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PLOS ONE*, 15(2), e0228987. https://doi.org/10.1371/journal.pone.0228987

Feldman, B., Martin, E. & Skotnes, T. (2012). Big Data in Healthcare Hype and Hope. Haettu 10.4.2021 osoitteesta https://www.ghdonline.org/uploads/Big-data-in-healthcare_B_Kaplan_2012.pdf

Fredriksson, C., Mubarak, F., Tuohimaa, M., & Zhan, M. (2017). Big Data in the Public Sector：A Systematic Literature Review. *Scandinavian Journal of Public Administration*, *21*(3), 1–23

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big Data concepts, methods, and analytics. *International journal of information management*, 35(2), 137-144.

Gantz, J. & Reinsel, J. (2011). Extracting Value from Chaos. *IDC's Digital Universe Study*

Gartner. (2021). Big Data. Noudettu 25.4.2021 osoitteesta https://www.gartner.com/en/information-technology/glossary/Big-data

Gupta, D. & Rani, R. (2019). A study of Big Data evolution and research challenges. *Journal of Information Science, 45*(3), 322-340. doi:10.1177/0165551518789880

Hariri, R. H., Fredericks, E. M. & Bowers, K. M. (2019). Uncertainty in Big Data analytics: Survey, opportunities, and challenges. *Journal of Big Data, 6*(1), 1-16. doi:10.1186/s40537-019-0206-3

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A. & Ullah Khan, S. (2015). The rise of "Big Data" on cloud computing: Review and open research issues. *Information Systems (Oxford), 47*, 98-115. doi:10.1016/j.is.2014.07.006

Hsieh, H.-F., & Shannon, S. (2005). Three Approaches to Qualitative Content Analysis. *Qualitative Health Researc*h, 15(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hu, M. (2015). Small Data Surveillance v. Big Data Cybersurveillance. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2731344

Hussain, K. & Prieto, E. (2016) Big Data in the Finance and Insurance Sectors. Teoksessa Cavanillas, J. ym., (toim.), New Horizons for a Data-Driven Economy. *Springer International Publishing* https://doi.org/10.1007/978-3-319-21569-3_12

Ishwarappa & Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and hadoop technology. *Procedia Computer Science, 48*, 319-324. doi:10.1016/j.procs.2015.04.188

Kaisler, S., Armour, F., Espinosa, J. A. & Money, W. (2013). Big Data: Issues and challenges moving forward. (s. 995-1004) *Institute of Electrical and Electronics Engineers (IEEE*). doi:10.1109/hicss.2013.645

Katal, A., Wazid, M., & Goudar, R. H. (2013). Big Data: Issues, challenges, tools and Good practices. *2013 Sixth International Conference on Contemporary Computing (IC3)*. https://doi.org/10.1109/ic3.2013.6612229

Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. & Salehian, S. (2018). The 10 vs, issues and challenges of Big Data. (s. 52-56) ACM. doi:10.1145/3206157.3206166

Kitchin, R. & McArdle, G. (2016). What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1), 205395171663113. doi:10.1177/2053951716631130

Lukoianova, T. & Rubin, V. L. (2014). Veracity roadmap: Is Big Data objective, truthful and credible? *Advances in Classification Research Online, 24*(1), 4. doi:10.7152/acro.v24i1.14671

Mahrt, M. & Scharkow, M. (2013). The value of Big Data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20-33. doi:10.1080/08838151.2012.761700

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburth, C. & Hung Byers, A. (2011*).* Big Data: The next frontier for innovation, competition, and productivity.

Mayer-Schönberger, V. & Cukier, K. (2013). Big Data: A revolution that will transform how we live, work and think.

McAfee, A., Brynjolfsson, E., Davenport, T., Patil, D., & Barton, D. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 90(10), 60-68.

McKinsey. (2011). Big Data: The Next frontier for innovation, competition, and productivity. Haettu 13.3.2021 osoitteesta https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/big-data-the-next-frontier-for-innovation

Mikalef, P., Boura, M., Lekakos, G. & Krogstie, J. (2019). Big Data analytics and firm performance: Findings from a mixed-method approach. *Journal of Business Research, 98*, 261-276. doi:10.1016/j.jbusres.2019.01.044

Nasser, T. & Tariq, R. (2015). Big Data, Big challenges. Journal of Computer *Engineering & Information Technology, 9307, 2.*

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: what can be done? Assessment & Evaluation in Higher Education, 33(3), 301–314. https://doi.org/10.1080/02602930701293231

Opresnik, D. & Taisch, M. (2015). The value of Big Data in servitization. *International Journal of Production Economics, 165*, 174-184. doi:10.1016/j.ijpe.2014.12.036

Oracle. (2021) What is Big Data? Noudettu 10.4.2021 osoitteesta https://www.oracle.com/Big-data/what-is-Big-data/

Panimalar, M. S. (2019). Survey paper on Big Data based isolation security by smartcard authentication system. *International Journal for Research in Applied Science and Engineering Technology, 7*(3), 846–853. doi:10.22214/ijraset.2019.3148

Pappas, I. O., Mikalef, P., Giannakos, M. N., Krogstie, J., & Lekakos, G. (2018). Big Data and business analytics ecosystems: paving the way towards digital transformation and sustainable societies. *Information Systems and E-Business Management*, 16(3), 479–491. https://doi.org/10.1007/s10257-018-0377-z

Rehman, M. H. u., Chang, V., Batool, A. & Wah, T. Y. (2016). Big Data reduction framework for value creation in sustainable enterprises. *International Journal of Information Management, 36*(6), 917-928. doi:10.1016/j.ijinfomgt.2016.05.013

Seddon, J. & Currie, W. (2016). A model for unpacking Big Data analytics in high-frequency trading. *Journal of Business Research, 70* doi:10.1016/j.jbusres.2016.08.003

Shafer, T. (2017). The 42 V's of Big Data and data science. Haettu 9.5.2021 osoitteesta https://www.kdnuggets.com/2017/04/42-vs-Big-data-data-science.html

Sharma, V., Pandey, B., & Kumar, V. (2016). Importance of Big Data in financial fraud detection. International Journal of Automation and Logistics, 2(4), 332. https://doi.org/10.1504/ijal.2016.080339

Sivarajah, U., Kamal, M. M., Irani, Z.,& Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research,* 70, 263-286.

Southerton, C. (2020). Datafication. Teoksessa L. A. Schintler & C. L. McNeely (toim.), Encyclopedia of Big Data (s. 1-4). Cham: Springer International Publishing. doi:10.1007/978-3-319-32001-4_332-1

Sun, Z., Strang, K. & Rongping, L. (2018). 10 Bigs: Big Data and its ten Big characteristics. BAIS No. 17010, PNG University of Technology.

Templier, M., & Paré, G. (2015). A Framework for Guiding and Evaluating Literature Reviews. *Communications of the Association for Information Systems,* 37, pp-pp. https://doi.org/10.17705/1CAIS.03706

Ularu, E., Puican, F., Apostu, A. & Velicanu, M. (2012). Perspectives on Big Data and Big Data analytics. *Database Systems Journal.*

Waller, M. A. & Fawcett, S. E. (2013). Data science, predictive analytics, and Big Data: A revolution that will transform supply chain design and management. *Journal of Business Logistics, 34*(2), 77–84. doi:10.1111/jbl.12010

Yaseen, H. K. & Obaid, A. M. (2020). Big Data: Definition, architecture & applications. JOIV : *International Journal on Informatics Visualization*, 4(1), 45-51. doi:10.30630/joiv.4.1.292

Özköse, H., Arı, E. S. & Gencer, C. (2015). Yesterday, today and tomorrow of Big Data. *Procedia, Social and Behavioral Sciences, 195*, 1042–1050. doi: 10.1016/j.sbspro.2015.06.147

Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of Big Data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, 69(5), 1562–1566. https://doi.org/10.1016/j.jbusres.2015.10.017