

Susanna Juurinen

Big data -teknologian perusteet ja mahdollisuudet

Metropolia Ammattikorkeakoulu

Insinööri (AMK)

Tuotantotalous

Insinöörityö

10.4.2013

Tekijä Otsikko	Susanna Juurinen Big data -teknologian perusteet ja mahdollisuudet
Sivumäärä Aika	41 sivua + 1 liite 10.4.2013
Tutkinto	insinööri (AMK)
Koulutusohjelma	tuotantotalous
Ohjaajat	lehtori Sakari Lind ryhmäpäällikkö Pasi Kolehmainen
<p>Työn tarkoituksena oli tutkia tiedonhallintaa ja big data -ratkaisujen käytännön toteutusta, mahdollisuuksia, haasteita ja hyötyjä. Työssä kuvataan tiedon määrän kehitystä, keräysprosesseja ja -menetelmiä sekä pohditaan tiedonhallinnan vaatimuksia ja tietoturvallisuutta. Työn tarkoitus oli selvittää tiedonhallinnan ja analytiikan uusimman trendin, big datan, vaikutusmahdollisuuksia organisaatioiden toimintaan.</p> <p>Tiedon analytiikkaa voidaan valjastaa liiketoiminnan tai muiden toimialueiden tarpeisiin tutkimalla esimerkiksi asiakaskäyttäytymistä erilaisissa tilanteissa ja ympäristöissä. Tämä tuo yritykselle aivan uudentyyppisiä toimintamahdollisuuksia, kuten yksilöidyn ja kohdennetun markkinoinnin. Joskus toimintaa ja tapahtumia on tarve tutkia pitkältikin ajanjaksolta vertaillen useamman tietolähteen sisältöä keskenään. Väärinkäyttötapauksissa taas lähdejärjestelmään tallennettujen tietojen eheys saattaa joutua kyseenalaiseksi. Näistä syistä, tai esimerkiksi PCI-standardin velvoittamana, tarve tapahtumatiedon keskitys-, hallinta- ja analysointiratkaisuille kasvaa jatkuvasti. Yhä useammin myös saatavilla olevan informaation määrä on liian suuri hallittavaksi perinteisillä analysointiratkaisuille. Big data -teknologialla monimuotoisen ja määrältään suuren informaatiojoukon käsittely ja analysointi voidaan suorittaa tehokkaasti ja nopeasti.</p> <p>Työ toteutettiin keräämällä tietoa tutkimuslaitosten julkaisuista, lehtiartikkeleista ja raportoiduista käyttökohde-esimerkeistä. Toimialakohtaisiin esimerkkeihin saatiin lisänäkemystä myös asiantuntijahaastatteluilla. Työ soveltuu yleiskuvaukseksi ja perusmateriaaliksi niille henkilöille, jotka työskentelevät tiedonhallinnan parissa tai kehittävät analyttisiä ratkaisuja.</p>	
Avainsanat	Big data, lokien hallinta, tiedon analysointi, tietoturvallisuus

Author Title	Susanna Juurinen Basics and the potential of big data technology
Number of Pages Date	41 pages + 1 appendices 10 April 2013
Degree	Bachelor of Engineering
Degree Programme	Industrial Management
Instructors	Sakari Lind, Senior Lecturer Pasi Kolehmainen, Group Manager
<p>The goal of this thesis was to study data management and big data solutions and to examine their opportunities, challenges and benefits. The thesis describes the evolution of information along with the processes and technologies of collecting information. Also data integrity and requirements were considered. This thesis investigates the effect which the newest business trend, big data, has on business and other activities.</p> <p>Data is used to improve business and other field activities such as analysis of customer behavior in different situations and environments. This brings up new opportunities like individualized and targeted marketing. In some cases actions and events need to be examined on a long term basis and data from different sources to be compared with each other. In case of misuse, there is the risk of the original information being corrupted. For these reasons or, for example, due to PCI compliance, the need to centralize, manage and analyze logs and other information is growing rapidly. More often the volume of data is too large to be handled with traditional analysis methods. With big data technology the analysis results of large amount of diverse data can be achieved faster and more efficiently.</p> <p>This thesis was carried out by collecting information from publications of research institutes, articles in the media and from reported cases of big data utilization. Also interviews with specialists were used to add perspective to the case studies. This thesis can serve as an outline and introduction to people, who work with data management or develop analytical solutions.</p>	
Keywords	Big data, log management, data analysis, data security

Sisällys

Lyhenteet ja käsitteet

1	Johdanto	1
2	Tiedonhallinnan kehityskaari lokeista big dataan	2
3	Tiedon määrän kasvu ja tehokkaan käsittelyn potentiaali	5
4	Big data -alustateknologia	9
4.1	Käsittelyn menetelmät	10
4.1.1	Rinnakkaisprosessointi	10
4.1.2	Muistiprosessointi	12
4.1.3	Virtausprosessointi	13
4.2	Tallennusympäristö	14
4.3	Pilviteknologia	15
5	Tiedonkeruu ja avustavia teknologioita	16
5.1	Kaupallinen tieto	17
5.2	Laitteiden välinen langaton koneviestintä	18
5.3	Älykäs sähköverkko	20
5.4	Mittautustietojen hallintajärjestelmät	21
6	Tiedon ja sen analysoinnin vaatimukset	22
7	Sovellusalueita ja esimerkkejä	24
7.1	Liiketoiminnan tehostaminen	25
7.2	Markkinointi	26
7.3	Energiankulutus	27
7.4	Terveystieteiden huolto	29
7.5	Bioteknikka	30
7.6	Kiinteistöala	31
7.7	Teollisuus	32
8	Mahdollisuudet ja haasteet	34
9	Yhteenveto	40
	Lähteet	42

Liitteet

Liite 1. Esimerkki Windows-järjestelmän kirjautumislukista epäonnistuneen kirjautumisyhteyden tapauksessa.

Lyhenteet ja käsitteet

3G	Lyhenne kolmannen sukupolven matkapuhelinteknologioille.
Ad hoc	Langattomien lähiverkkojen välinen tukiasematon yhteystapa.
BI	<i>Business Intelligence</i> . Liiketoiminnan analysointijärjestelmä.
BSON	<i>Binary JSON</i> . MongoDB:n verkossa ja tallennustilassa käytettävä tiedon-siirtoformaatti.
Genomi	Organismin koko DNA:han koodattu perintöaines.
Hadroni	Vahvoihin vuorovaikutuksiin osallistuva alkeishiukkanen.
HBase	Apachen Javalla ohjelmoima jaettu tietokantajärjestelmä.
HDFS	<i>Hadoop Distributed File System</i> . Javalla Hadoop-ohjelmistokehystä var-ten ohjelmoitu jaettu ja skaalautuva tiedostojärjestelmä.
HIPAA	<i>Health Insurance Portability and Accountability Act</i> . Yhdysvalloissa ter-veydenhuoltoa koskeva yksityisyysmääräys.
Hive	Apachen Javalla ohjelmoima tietovarastojen hallintaratkaisu.
Java	Laitteistoriippumaton oliopohjainen ohjelmointikieli.
JSON	<i>JavaScript Object Notation</i> . Yksinkertainen tiedonsiirtomuoto, jota voi-daan käyttää mm. JavaScript-ohjelmissa.
Klusteri	Joukko laitteita, jotka ovat verkotettuja keskenään hoitamaan yhteisiä tehtäviä.
M2M	<i>Machine to machine</i> . Langattomaan koneiden väliseen viestintään viittaa-va termi.
MDMS	<i>Meter Data Management System</i> . Mittaustietojen hallintajärjestelmä.

Noodi	Laite, joka on osa klusteria.
NoSQL	Tietokantamalli, jolla on suurempi skaalautuvuus ja suorituskyky kuin perinteisellä relaatiotietokannalla.
OI	<i>Operational Intelligence</i> . Toimintojen analysointijärjestelmä.
PCI	<i>Payment Card Industry</i> . PCI-standardi sisältää luottokorttiyhtiöiden määrittelemät tietoturvan vähimmäisvaatimukset.
RAM	<i>Random Access Memory</i> . Tietokoneen keskusmuisti.
Replikointi	Tiedon monistus.
RFID	<i>Radio Frequency Identification</i> . Radiotaajuustunnistus.
SEM	<i>Security Event Management</i> . Tietoturvatapahtumien hallinta.
SIEM	<i>Security Information and Event Management</i> . Tietoturvainformaation ja tapahtumien hallinta.
SIM	<i>Security Information Management</i> . Tietoturvainformaation hallinta.

1 Johdanto

Lähes kaikki tehdyt toimenpiteet, ympäristössä tapahtuvat muutokset ja laitteen tai järjestelmän tapahtumat jättävät elektronisen jäljen. Yhä useammin käytettävät työkalut, elämää helpottavat ratkaisut tai yhteiskuntaa tukevat rakenteet hyödyntävät modernia teknologiaa. Ympäristön muutoksia pyritään mittaamaan, tallentamaan ja analysoimaan erilaisilla tavoilla oikean kehityssuunnan takaamiseksi niin liiketoiminnassa kuin perinteisissä arkiaskareissakin. Jo siitä saakka, kun tapahtumia ja ilmiöitä on osattu kirjata ylös, on pyritty tallentamaan mm. kaupankäyntiin tai taivaankappaleiden liikkeisiin liittyvät oleelliset asiat, jotta niiden perusteella on voitu kehittää toimintatapoja ja tehostaa resurssien käyttöä pitkällä aikajänteellä. Moderni teknologia on tuonut markkinoille menetelmiä, joilla toiminnan ja muutoksien analytiikkaa voidaan nostaa aivan uudelle tasolle. Näiden menetelmien avulla ei enää ainoastaan voida tehostaa toimintaa tai ottaa opiksi virheistä, vaan jopa ennustaa tulevaa ja varautua todennäköisiin muutoksiin entistä tehokkaammin ja tarkemmin.

Tapahtumatiedon keruulla, hallinnalla ja analysoinnilla on ollut useita eri käyttötarkoituksia ja konseptista on käytetty useita eri termejä, joista uusimpana big data. Big data -käsitteellä tarkoitetaan sellaisia tietojoukkoja, joiden koko on suurempi kuin mitä perinteisillä tietokantaratkaisuilla voidaan kerätä, tallentaa, hallita ja analysoida. Tämä tietojoukkojen kirjo pitää sisällään merkittävän määrän sellaista asiakas-, toimittaja- ja toimintatietoa, jota aiemmin ei ole pidetty merkittävänä. Tietoa saadaan erilaisista tunnistimista ja antureista, kuten matkapuhelimista, älykkäistä energiamittareista, ajoneuvoista tai teollisuuden laitteista, ja niistä kerättyä tietoa kutsutaan myös asioiden internetiksi [1, s. 1]. Digitaalisessa maailmassa tietoa generoituu kaikesta toiminnasta aina yksilön kommunikoinnista ja ostotottumuksista sään vaihteluun, energian tuottamiseen ja astronomisiin muutoksiin saakka.

Insinööriyön tavoite

Työn tavoitteena on tutustua uusimpaan ohjaavan analytiikan ratkaisuun big dataan, sen hyödynnettävyyteen ja skaalautuvuuteen yhteiskunnan eri sektoreilla. Työssä tutkitaan, miten olemassa olevaa tietopohjaa ja sen analysointia voitaisiin laajentaa koskemaan uusiakin tietolähteitä ja kuinka jo olemassa olevia tietolähteitä pystyttäisiin hyödyntämään analyyseissä uudella tavalla. Esille tuodaan myös uusia teknologiaratkaisuu-

ja, joita voidaan valjastaa tiedon hyödyntämisen tarpeisiin. Lisäksi työssä pohditaan kehityssuunnan mukanaan tuomia haasteita ja mahdollisuuksia sekä yksilön että yhteiskunnan näkökulmasta. Työssä on käytetty lähteinä aihepiirin uusimpia julkaisuja, asiantuntijahaastatteluita ja aihealueeseen liittyviä tutkimuksia sekä käytännön toteutuksia. Työn toteutukseen vaikuttaa merkittävästi kirjoittajan useiden vuosien työhistoria lokihallinnan, valvonnan ja niihin liittyvien sovellusalueiden parissa. Suomenkielistä materiaalia on niukasti saatavilla, mistä johtuen kuvamateriaalin esityskielenä on pääsääntöisesti englanti.

Insinööriyön rakenne ja rajaukset

Insinööriyö koostuu käytännössä neljästä eri osakokonaisuudesta. Ensimmäisenä tutustutaan tietolähteisiin, tiedon tallennukseen ja big data -teknologiaan yleisellä tasolla. Toinen osakokonaisuus keskittyy analytiikan perusteisiin ja käytettävän tiedon laadullisiin vaatimuksiin sekä haasteisiin. Kolmannessa osiossa tutustutaan tallennetun informaation analysoinnin ja muun hyödyntämisen sovellusalueisiin ja viimeisessä osiossa pohditaan tiedon käytön mahdollisuuksia sekä haasteita. Työssä sivutaan teknisiä ratkaisuja, mutta ei oteta kantaa markkinoilla oleviin kaupallisiin toteutuksiin tai mennä teknisten ratkaisujen yksityiskohtiin. Sovellusalueita on huomattava määrä ja uusia kehitetään jatkuvasti. Tässä työssä keskitytään kokonaisuuksiin, joiden merkitys on pääsääntöisesti universaali tai muuten yhteiskunnallisesti merkittävä.

2 Tiedonhallinnan kehityskaari lokeista big dataan

Alun perin lokeilla on tarkoitettu laivoissa ylläpidettävää lokikirjaa, johon on kirjattu navigointia helpottavia ja matkantekoon liittyviä asioita. Tietotekniikassa lokeilla viitataan automaattisesti muodostuviin järjestelmien keräämiin merkintöihin, joista on todennettavissa järjestelmän tapahtumat, siellä tehdyt toimenpiteet, virheet ja muutokset. Lokien avulla on mahdollista esimerkiksi todentaa onnistuneet järjestelmäpäivitykset, selvittää järjestelmässä tapahtuneita ongelmatilanteita, tutkia väärinkäyttötilanteita tai tapahtumaketjuja. Tähän saakka lokien käsittely onkin ollut järjestelmävalvonnan ja ongelman selvityksen keskiössä, sillä niistä on voitu saada vastaukset kysymyksiin kuka, mitä, milloin, miten ja missä.

Järjestelmien pitäminen toimintakunnossa on perinteisesti ollut yksi tärkeimmistä kokonaisuuksista tietotekniikan aikakaudella, oli sitten kyse yksityisestä liiketoiminnasta, julkishallinnon ylläpitämisestä tietovarastoista tai käyttäjän henkilökohtaisesta tietokoneesta. Tietotekniikan käytön yleistyminen ja käytettävien järjestelmäympäristöjen kasvu johtivat siihen, ettei pelkkä lokien lukeminen ongelmatilanteiden analysoinnissa enää riittänyt, vaan automaattisen seurannan tarve kasvoi. Jo useita vuosia järjestelmiä on valvottu lokeja seuraamalla ja niitä tulkitsemalla sekä manuaalisesti että automaattisesti. Lokien tulkinta onkin automaattisten valvontajärjestelmien tärkeimpiä rooleja edelleen. Valvontajärjestelmiin pystytään määrittelemään etukäteen esimerkiksi viesteissä ilmenevien merkkijonojen tai tapahtumaa muuten identifioivien ominaisuuksien perusteella raja-arvoja ja parametreja, joiden toteutuminen laukaisee automaattisen hälytyksen tai jopa automaattisia korjaustoimenpiteitä. Tällaisista tapahtumista mainittakoon esimerkiksi tietoliikenneyhteyksien katkeaminen tiettyyn resurssiin, palvelun epäonnistunut käynnistäminen, epäonnistuneet kirjautumisyritykset tai tietyllä tunnuksella kirjautuminen. Esimerkki Windows-palvelimelle tallentuneesta lokimerkinnästä tapauksessa, jossa käyttäjä on pyrkinyt tunnistautumaan käyttäen väärää salasanaa, on liitteessä 1.

Modernit valvontajärjestelmät hyödyntävät lokien lisäksi myös muuta laitteista ja järjestelmistä saatavaa tietoa, kuten laitteiden metristä tietoa. Niiden avulla on mahdollista tunnistaa tilanteita, joissa tekniset resurssit saavuttavat ennakoimatta ennalta määritellyjä raja-arvoja. Metriikkaa ja kerättyä lokitietoa yhdistämällä voidaan hälytyksien lauetessa selvittää ongelmien tai kuormitustilojen aiheuttajat mahdollisesti jo ennen merkittävää haittaa toiminnalle. Lokivalvonta tulee lokitiedon laajemmasta hyödyntämisestä huolimatta säilyttämään asemansa osana yrityksen tietoturvastrategiaa jatkossakin. Tiedon laajemman hyötykäytön myötä saadaan kuitenkin useita lisänäkökulmia myös valvontaan. Perinteisesti lokeista on voitu vastata kysymyksiin kuka, mitä, milloin, miten ja missä, mutta hyödyntämällä koko tietokapasiteettia voimme parantaa kykyä vastata myös kysymykseen miksi. Jos löytyy menetelmä, algoritmi tai tapahtumayhdistelmä, jonka avulla pystytään selvittämään juurisyy koko tietyn tapahtumaketjun syntymiselle, pystytään entistä paremmin turvaamaan organisaation tai yksilön tietoturvasuus jatkossa.

Tietotekniikan käytön yleistyttyä myös siihen kohdistuvien uhkatekijöiden määrä on kasvanut. Yhä useammat tekniset ratkaisut suojelevat olemassa olevaa tietojärjestelmäinfrastruktuuria ja sen sisältämää tietopääomaa. Järjestelmissä itsessäänkin ilmenee joskus tietoturva-aukkoja, joiden löytäminen ja tukkiminen ovat osa normaalia yllä-

pito- ja päivitysprosessia. Lokihallinnan näkökulmasta tietoturvallisuutta on pyritty parantamaan hyödyntämällä SIEM-konseptia eli tietoturvainformaation ja -tapahtumien hallintaa. Siinä keskitytään analysoimaan erityisesti tietoturvaan liittyviä lokimerkintöjä ja laitteiden metriikkaa uhkatilanteiden ja riskien tunnistamiseksi. Analyysieihin yhdistetään organisaation ulkopuolella havaittuja uhkia ja tunnistettuja heikkouksia, jotta tietoturvallisuutta voidaan parantaa myös ennakoivalla toiminnalla.

Perinteisesti ongelmia ja uhkia on etsitty kuin neulaa heinäsuovasta, mutta vaihtamalla näkökulma siihen, että poistetaan heinät neulan ympäriltä, voidaan organisaatiolle rakentaa tarvittava turvaverkko entistä kattavammin ja ennakoivammin. Huomioimalla myös sellaiset tietolähteet, joita on perinteisesti pidetty merkityksettöminä tai joiden olemassaoloa ei ole edes huomattu, voidaan lyhentää tietorikollisen toiminta-aikaa useista päivistä tai viikoista tunteihin tai jopa minuutteihin. SIEM on yhdistelmä kahdesta edeltäjästään (SEM ja SIM), joissa tietoturvaan liittyvän informaation ja tietoturvaan liittyvien tapahtumien hallinta ovat olleet erillään [2].

SIEM-ratkaisuja on markkinoilla useita. Tammikuussa 2012 itsenäinen tietoturvaohjelmistoja vertaileva organisaatio, Mosaic Security Research, sai selvityksessään yksilöityä 85 kaupallista SIEM-järjestelmää, jotka tarjoavat reaaliaikaista tietoturvahälytyksien analysointia ja raportointia sekä laitteista että sovelluksista [2]. Näiden järjestelmien avulla voidaan kerätä keskitetysti kaikki lokidata järjestelmän levypinnalle, jotta kriittistä lokitietoa ei menetettäisi esimerkiksi laiterikko- tai tietomurtotilanteessa. Levypinnalla tarkoitetaan tässä yhteydessä järjestelmän käyttöön allokoitua tallennustilaa. SIEM-tuotteen avulla lokitiedoista kerättävä informaatio voidaan kääntää luettavaan muotoon, sillä joskus lokit ovat kryptisiä ja niiden tulkinta vaatii valmistajan ohjeistusta. Kerätystä tiedosta voidaan myös tuottaa raportteja, taulukoita tai graafisia esityksiä, joista poikkeustilanteet havaitaan helposti ja jotka ovat ei-tekni senkin henkilön ymmärrettävissä ja analysoitavissa. Raportteja voidaan tuottaa olemassa olevien turvallisuus-, auditointi- tai hallinnollisten prosessien vaatimuksien mukaisesti. SIEM-järjestelmään voidaan generoida automaattisia hälytyksiä, jotta korjaaviin toimenpiteisiin voidaan ryhtyä välittömästi. Lisäksi informaation pitkäaikainen tallennus mahdollistaa tapahtumien ja tapahtumaketjujen tutkimisen myös jälkikäteen, jos tutkittavat tapaukset eivät ole tulleet ilmi tapahtumahetkellä tai välittömästi sen jälkeen.

Liiketoiminnassa vastaavanlaista tiedonlouhintaa ja analysointia on suoritettu erilaisilla Business Intelligence -järjestelmillä, joiden tarkoituksena on ollut avustaa johdon strategista päätöksentekoa. BI-järjestelmissä informaatio kerätään pääasiassa yrityksen

sisäisistä järjestelmistä, kuten toiminnanohjaus- tai kommunikaatiojärjestelmästä, tai yrityksen ulkopuolelta kilpailijoista tai markkinoiden kehitymisestä. [3.]

Lokien kokonaisvaltaisen hyödyntämisen näkökulmasta tiedon analysointikonseptit, kuten SIEM tai BI, ovat olleet hyvä alkua ja varmasti tulevat olemaan kiinteä osa tiedonhallinnan ja käsittelyn sovellusalueita myös tulevaisuudessa. Viimeisimpien tietoteknisten ratkaisujen myötä tiedonhallinnan merkitys pelkästään ongelmanselvityksen ja liiketoiminta-analytiikan apuvälineenä kasvattaa kuitenkin mittasuhteitaan myös organisaatioiden hallinnon, terveydenhuollon, energiateollisuuden ja globaalin kansantalouden näkökulmasta. Tämä ilmiö tunnetaan nimellä big data.

Big data -käsitteessä ei ole kyse isoista tiedoista, vaan isosta tiedon volyymistä ja sen moninaisuudesta. Big datassa lokilähteitä eivät ole enää pelkästään tietotekniset laitteet tai sovellukset, vaan analysoitavaa tapahtuma- tai tilatietoa voidaan kerätä useilla eri menetelmillä lähes mistä tahansa. Tietolähteinä voivat toimia mm. RFID-tunnisteet eli radiotaajuustunnisteet, meteorologiset anturit, genomi-informaatio, sosiaalisen median syötteet, hakuohjelmiin syötetyt hakusanat, internet-klikkaukset, kannettavien laitteiden paikkatiedot tai vaikka kanta-asiakaskortille kirjatut ostokset. Suurimmat haasteet liittyvät kerättävän tiedon hallintaan ja siihen, missä käyttötarkoituksissa tietoa voidaan hyödyntää. Useista eri lähteistä kerätty tieto on kompleksista ja vie huomattavasti tallennustilaa. Tiedon siistimistä, käsittelyä tai jakamista ei voida enää suorittaa tehokkaasti käytössä olevien teknologioiden avulla. Tästä syystä perinteiset tallennus-, haku- ja analysointimenetelmät eivät palvele big data -konseptia, vaan tiedonhallintaa varten täytyy löytää täysin uuden näkökulman ratkaisuja.

3 Tiedon määrän kasvu ja tehokkaan käsittelyn potentiaali

Tietoinfrastruktuuriratkaisujen toimittajan EMC:n toimeksiannosta vuonna 2010 tehdys- sä tutkimuksessa digitaalisen tiedon määrän arvioidaan nelinkymmenkertaistuvan vo- teen 2020 mennessä. Kuvassa 1 on kuvattuna tiedon määrän kasvu eksatavuina vo- desta 2005 vuoteen 2015. Rajusti kasvava tietomäärä on haaste, sillä tallennustila on rajallista ja sen lisäksi osa tiedosta vaatii erityishuomiota tietosuojanäkökulmasta. Tut- kimuksen mukaan vain pientä osaa tiedosta hyödynnetään, sillä vain noin 0,5 prosent- tia tiedosta päätyy analyysiin, vaikka tällä hetkellä olisi mahdollista hyödyntää jopa 23 prosenttia tallennetusta tiedosta big data -analyyseissä. [4]



Kuva 1. Digitaalisen tiedon määrän kasvuennuste [4].

Perinteisillä menetelmillä kohtuullisessa ajassa prosessoitavat tietokokonaisuudet eivät ole nykyvaatimuksiin riittäviä, mikä on rajoittanut huomattavasti suurien tietokokonaisuuksien käsittelyä esimerkiksi genomiteknologiassa tai monimutkaisissa fysiikan simulaatioissa. Suurien tietomäärien käsittelykyky on lähes välttämättömyys informaatiolähteiden moninkertaistuessa ja teknologian kehittyessä. Esimerkiksi maailman suurin hiukkaskiihdytin sisältää 150 miljoonaa sensoria, jotka tuottavat tietoa 40 miljoonaa kertaa sekunnissa. Kiihdyttimessä tapahtuu lähes 600 miljoonaa hadronitörmäystä sekunnissa, joista suodatuksen jälkeen jäljelle jää noin sata tutkijoihin kiinnostavaa tapahtumaa sekunnissa. Jos suodatusta ei tehtäisi, tiedonmäärä olisi valtava, noin 500 triljoonaa tavua päivässä, mikä on 200-kertainen määrä tietoa verrattuna kaikkiin muihin maailman tietolähteisiin yhteensä.

Ihmisgenomin dekoodaaminen vei aiemmin aikaa 10 vuotta, mutta nykyään se voidaan tehdä viikossa. Yhdysvaltalainen kauppaketju Walmart käsittelee yli miljoona asiakastapahtumaa joka tunti, jotka kaikki tallennetaan tietokantoihin. Sen tietokannoissa arvioidaan olevan yli 2,5 petatavua informaatiota, joka on 167 kertaa enemmän kuin Yhdysvaltain kongressin kirjastossa. Teknologisesti tallennetun tiedon määrä henkilöä kohden on arviolta kaksinkertaistunut 40 kuukauden välein 80-luvun alusta ja vuonna 2012 joka päivä syntyi noin 2,5 triljoonaa tavua digitaalista tietoa. Big data -konseptissa käsiteltävien tietokokonaisuuksien koko on suurempi kuin perinteisillä tietokanta- tai

ohjelmistoratkaisuilla voidaan tehokkaasti käsitellä. Tallennettujen tietokokonaisuuksien koon jatkuva kasvaminen asettaa luonnollisesti kehitysvaatimuksia myös big data -teknologioille. [5.]

Big data -teknologia on vaikuttanut huomattavasti myös tiedonhallinnan ammattilaisten kysyntään. Vuonna 2010 tiedonhallintaan ja analysointiin erikoistuneiden yritysten arvo oli yli 100 miljardia dollaria kasvaen vuosittain lähes 10 prosenttia eli kaksinkertaisesti muuhun ohjelmistoliiketoimintaan verrattuna [4]. Kansainvälinen ICT-tutkimus- ja konsultointiyritys Gartner on ennustanut, että big datan myötä tiedon käsittelyyn liittyvien työpaikkojen määrä tulee kasvamaan huomattavasti. Arvion mukaan vuoteen 2015 mennessä sen käsittely tarvitsee globaalisti jopa 4,4 miljoonaa uutta osaajaa ja kerrannaisvaikutuksen myötä jokaista IT-asiantuntijaa kohden muodostuu kolme uutta työpaikkaa muille toimialoille. [6.]

On laskettu, että big data -konseptin hyödyntäminen toisi kymmenen vuoden sisällä mm. Euroopan julkisen sektorin hallinnolle 150–300 miljardin euron vuosittaiset säästöt ja 300 miljardin dollarin säästöt Yhdysvaltojen terveydenhuoltosektorille [1, s. 62; 1, s. 50]. Euroopan julkisen sektorin potentiaaliset säästökohteet voidaan jaotella kolmeen osa-alueeseen: toiminnan tehostamiseen, petosten ja virheiden vähentämiseen ja verotusasteen nostamiseen. Taulukossa 1 on jaoteltu arvioidut säästöt osa-alueiden mukaan ja siinä on havaittavissa suurimman osan säästöistä tulevan nimenomaan toiminnan tehostamisen sektorilta.

Taulukko 1. Big data -analytiikan laskennalliset hyödyt Euroopan julkisen sektorin hallinnolle [1, s. 62].

		Kokonais- kustannukset Miljardia €	×	Tunnistettavia %	×	Säästöt %	=	Kokonaisarvo Miljardia €
Toiminnan tehostamisen säästöt	Toiminnan kustannukset	4,000		20–25		15–20		120–200
Virheiden ja petoksien vähentäminen	Tulonsiirto	2,500		1–3		30–40		7–30
Veronkeruun tehostaminen	Verokertymä	5,400		5–10		10–20		25–110
								150–300+

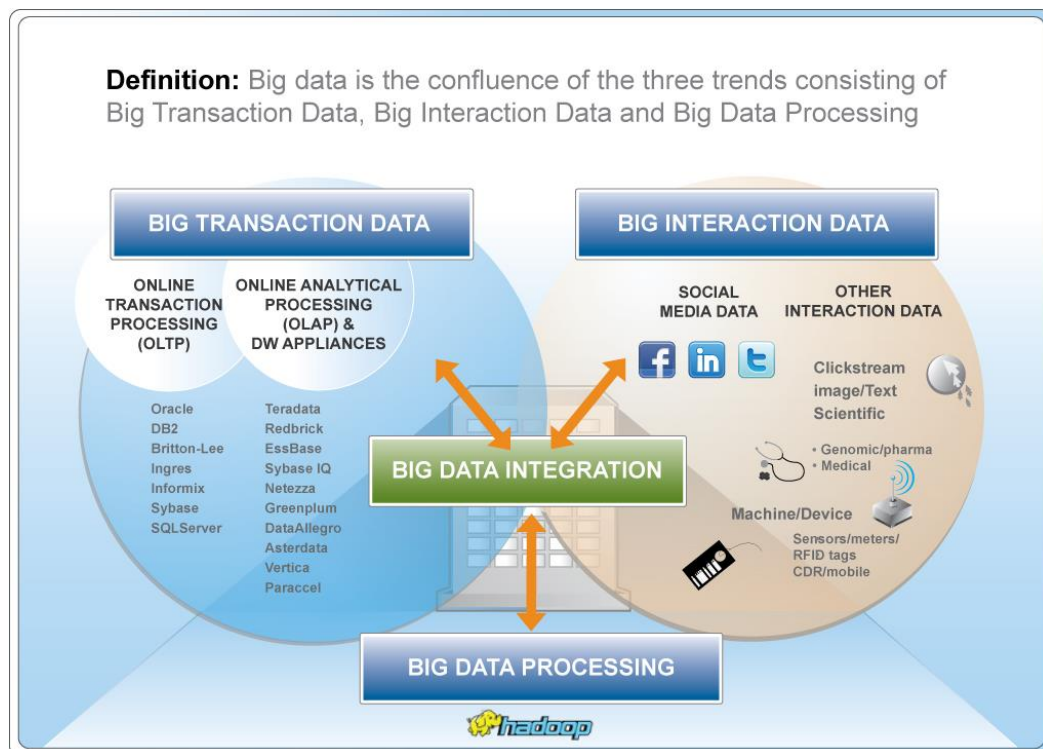
Toiminnan tehostamisen saralla arviot perustuvat siihen, että tunnistetut tehokkuustekijät koskevat 20–25 prosenttia toimintabudjetista ja potentiaaliset säästöt kattavat 15–20 prosenttia kyseisestä alueesta. Petos- ja virhetilanteissa säästöarviot ovat muodostuneet sillä perusteella, että epäselvyyksiä ilmenee 1–3 prosentissa tulonsiirtotapahtumia ja niiden aiheuttamia kustannuksia on mahdollisuus pienentää seurannalla ja analysoinnilla jopa 40 prosenttia. Verokuilun taas arvioidaan olevan noin 5–10 prosenttia Euroopan tasolla ja tästä 10–20 prosenttia olisi mahdollista pidättää hyödyntämällä big data -ratkaisuja. Edellä mainittujen lisäksi big data -analyysillä saavutetun tietopääoman avulla voidaan optimoida myös sellaisia toiminnallisia osa-alueita, jotka eivät suoraan tuo säästöjä, mutta tekevät toiminnasta tarkoituksenmukaisempaa. Tällaisiin voidaan lukea esimerkiksi korkealaatuiset palvelut, resurssien allokointi ja parempi tiedotus. [1, s. 61–62.]

Yhdysvaltojen terveydenhuoltosektorilla big data -ratkaisulla voidaan parantaa merkittävästi tuottavuutta, hoidon laatua, potilaskokemuksia ja kilpailukykyä sekä luoda uusia liiketoimintamalleja ja palveluita. Johdon konsultointiyritys McKinsey & Companyn liiketoiminnan ja talouden tutkimusosaston MGI:n arvion mukaan kaksi kolmasosaa 300 miljardin dollarin säästöistä voitaisiin saavuttaa pienentämällä kansallisen terveydenhuollon kustannuksia kahdeksalla prosentilla. Säästöjen edellytyksenä on kuitenkin se, että terveydenhuollon organisaatiot ja poliitikot ottavat kollektiivisesti kantaa olemassa oleviin organisaatorakenteisiin ja muihin kehitystä hidastaviin tekijöihin. Tämän seura-

uksena potilaat saisivat pienemmillä kustannuksilla laadukkaampaa hoitoa ja laajemman terveydenhuollollisen tietopohjan, jonka avulla heistä tulisi tiedostavampia terveydenhuoltojärjestelmän käyttäjiä. [1, s. 49–50.]

4 Big data -alustateknologia

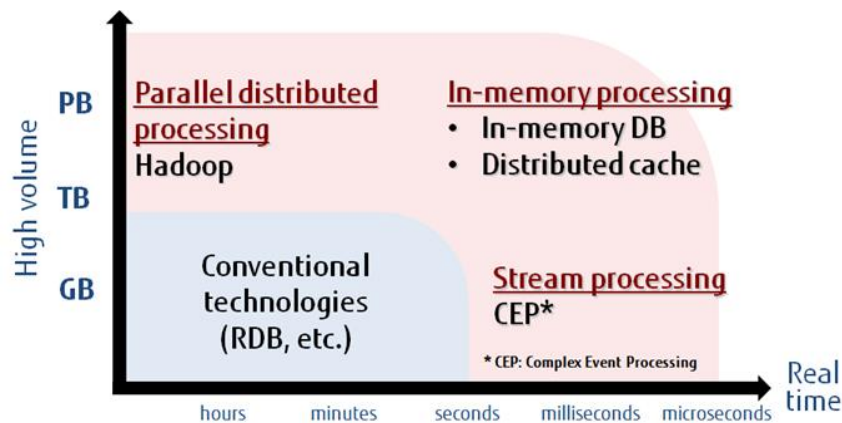
Peruselementeiltään big data -konsepti toimii kuten mikä tahansa analyysijärjestelmä. Tietoa tunnistetaan, sitä kerätään, tallennetaan, prosessoidaan ja lopputuotteena saadaan lisäarvoa tuottava yhteenveto kerätyistä tiedosta. Tietoa kerätään useista eri resursseista, kuten järjestelmien tuottamista tietokokonaisuuksista, jotka sijaitsevat usein joko lokitiedostoissa tai järjestelmien tietokannoissa. Tietoa voidaan kerätä myös esimerkiksi sosiaalisesta mediasta, mobiililaitteista tai erilaisista sensoreista. Kerätty informaatio kerätään big data -järjestelmään, jossa se prosessoidaan. Prosessoidusta tiedosta tuotetaan analyysejä erilaisilla analyysijärjestelmillä. Kuvassa 2 on kuvattu tiedon käsitteellinen kulku big data -järjestelmässä.



Kuva 2. Käsiteltävän informaation kerääminen [7].

4.1 Käsittelyn menetelmät

Tietoa voidaan prosessoida erilaisilla teknologia-alustoilla. Big data -ympäristössä tiedon prosessointia voidaan toteuttaa rinnakkaisesti suoritettuna tiedostojärjestelmässä, muistialustalla tai virtauslogiikalla. Oikean teknologian valinta riippuu siitä, kuinka reaaliaikaista prosessointia vaaditaan ja kuinka suuri on tiedon volyymi (kuva 3).



Kuva 3. Käsittelymenetelmät vaatimusten kasvaessa [8].

Seuraavana esitellään näiden prosessointiteknologioiden erityispiirteitä ja soveltuvuutta erityyppisiin prosessointivaatimuksiin.

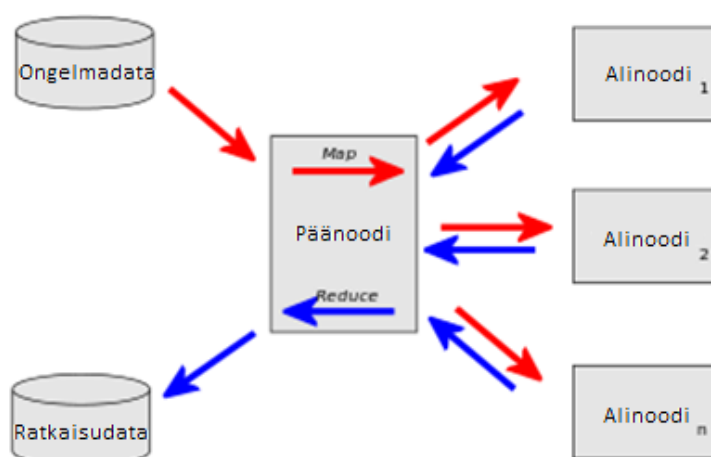
4.1.1 Rinnakkaisprosessointi

Tässä prosessointimallissa tietoa käsitellään klusterissa eli samaa tehtävää toteuttavassa laitejoukossa samanaikaisesti usealla eri noodilla eli klusterin laitteella. Tämä malli mahdollistaa sen, että suurien tietomäärien käsittely voidaan jakaa koko klusterin laajuudelle, jolloin tieto voidaan käsitellä todella nopeasti. Vaikka rinnakkaisprosessoinnissa pystytään hallitsemaan tehokkaasti hyvin suuria tietomääriä, analyysien tuottaminen ei ole yhtä nopeaa kuin jäljempänä esiteltävissä muisti- tai virtausprosessoinnissa. Jos analyysien tuottaminen mahdollisimman reaaliaikaisesti on erittäin kriittistä, tulisi harkita vaihtoehtoisia prosessointimenetelmiä.

Rinnakkaisprosessoinnin mallia käyttää esimerkiksi Javalla ohjelmoitu, avoimen lähdekoodin ohjelmistokehys Apache Hadoop, joka toimii useimpien big data -toteutuksien alustana. Hadoopissa laajat tietomassat tallennetaan hajautetusti klusteriin tiedon kä-

sittelyyn käytettävän ajan optimoimiseksi. Hadoop-teknologiassa sovellus jaetaan useisiin pieniin toimintaosiin, joista jokainen voidaan suorittaa klusterin jokaisella noodilla. Jako toteutetaan MapReduce-algoritmillä, jota Hadoopin MapReduce-moottori käyttää jakaakseen työt ympäri klusteria. Lisäksi Hadoop toimii jaettuna tiedostojärjestelmänä, joka tallentaa tietoa klusterin noodeille tarjoten runsaasti tiedonsiirtokaistaa klusterin sisällä. MapReduce ja tiedostojärjestelmä on suunniteltu siten, että ohjelmistokehitys pystyy automaattisesti selviytymään noodin rikkoutumisesta keskeyttämättä sovellusten toimintaa. Hadoop-alusta koostuu ytimeistä, MapReduce-algoritmista ja Hadoop-yhteensopivasta tiedostojärjestelmästä. Lisäksi siihen luetaan kuuluvaksi pienempiä ohjelmistototeutuksia, kuten tietovarastojen hallintaratkaisu Apache Hive ja jaettu tietokantajärjestelmä Apache HBase. [9.]

MapReduce on ohjelmointimalli, jonka avulla voidaan tuottaa rinnakkaisprosessointiohjelmia suurien tietomäärien käsittelemiseksi. Sen algoritmi sisältää yksinkertaiset Map- ja Reduce-funktiot, jotka keskittyvät ongelmalogiikkaan samalla, kun MapReduce-moottori huolehtii tehtävien kanavoinnista klusterille ja tiedonsiirrosta sekä kommunikoinnista järjestelmän osien välillä. Map-funktiolla saapuva informaatio hajautetaan alinoodille pienissä tehtäväosissa, jotka alinoodit käsittelevät ja palauttavat vastauksen päänoodille. Reduce-funktiossa päänoodi vastaanottaa alinoodien osavastaukset ja yhdistää ne tuottaakseen kokonaisvastauksen alkuperäiseen tehtävään. Kuvassa 4 on esiteltynä funktioiden toimintakaavio. [10.]



Kuva 4. Map- ja Reduce-funktiologiikka [10].

Jotta tehtävät voidaan aikatauluttaa tehokkaasti, jokaisen Hadoop-yhteensopivan tiedostojärjestelmän tulee tarjota paikkatietoa eli sen kytkimen ja kehikon nimi, johon noodit on liitetty. Sovellukset hyödyntävät tätä tietoa, jos sen käyttämällä noodilla ilmenee virhetilanne. Tällöin järjestelmä voi jatkaa työtä käyttämällä samaan kytkimeen liitettyä toista noodia välttääkseen runkoverkon turhaa kuormittamista. Esimerkiksi HDFS-tiedostojärjestelmä käyttää tätä menetelmää tiedon replikoinnissa eli tiedon monistuksessa, jolloin sama tieto hajautetaan eri kehikkoihin vikasietoisuuden parantamiseksi. HDFS suunniteltiin käsittelemään hyvin suuria tiedostoja. Se tallentaa suuret tiedostot usealle eri laitteelle ja replikoi tiedostot niin ikään usealle noodille. HDFS ei ole ainoa Hadoopin tukema tiedostojärjestelmä, vaan mitä tahansa jaettua tiedostojärjestelmää voidaan käyttää. [9.]

4.1.2 Muistiprosessointi

Muistiprosessointimallissa tieto tallennetaan levypinnan sijaan järjestelmän keskusmuistiin (RAM). Tämä mahdollistaa nopean tiedon käsittelyn, ja lisäksi tulokset ovat erittäin nopeasti luettavissa. Muistista lukeminen on jopa sata kertaa nopeampaa kuin levyltä luku, ja nykyaikaisilla työkaluilla voidaan kompressoida myös muistiin tallennettavaa tietoa. Muistiprosessoinnin haittana on käyttäjämäärien ja tiedon volyymin kasvu, joka edellyttää myös muistikapasiteetin kasvattamista. Tämä saattaa pian näkyä myös kasvaneina laitteistokustannuksina. Muistiprosessointia sopii harkita big data -analytiikassa silloin, kun analyysien tuottamisen tulee olla lähes reaaliaikaista ja hidasteet estäisivät käyttäjää tekemästä tärkeitä päätöksiä. Jos yritys tuottaa merkittävän määrän tallennettavaa tietoa, eikä raportointivaatimuksissa nopeutta ole priorisoitu, muistiprosessointi ei mallina ole välttämättä sopivin vaihtoehto. [11.]

Muistiprosessointimallia käytetään esimerkiksi avoimen lähdekoodin MongoDB-järjestelmässä. MongoDB on ohjelmistotalo 10genin kehittämä, C++ -ohjelmointikielellä kirjoitettu NoSQL-alusta. Siinä tieto tallennetaan perinteisen relaatiotietokannan sijaan yksinkertaiseen tiedonsiirtomuotoon JSON-dokumenteiksi, joilla on dynaaminen BSON-formaatin eli mongoDB:ssä käytettävän binäärisen tiedonsiirtoformaatin kaava. MongoDB perusominaisuuksiin kuuluu ad hoc -kyselyiden eli langattomien lähiverkkojen välisten kyselyiden ja JavaScript-ohjelmien tukeminen. [12.]

Mikä tahansa MongoDB-dokumentin kenttä voidaan indeksoida, ja myös sekundäärinen tason indeksointi on tuettu. Arkkitehtuuri perustuu niin sanottuun

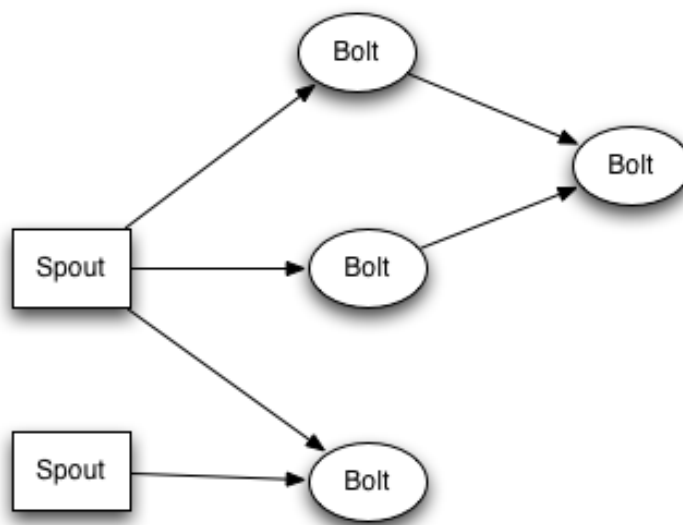
master-slave -malliin, jossa päätietokanta eli master lukee ja kirjoittaa saapuvaa tietoa. Alitietokanta eli slave kopioi tietoa master-instanssilta, mutta kykenee vain lukemaan tai varmistamaan tietoa. Jos päätietokanta lakkaa toimimasta, alitietokannat valitsevat automaattisesti uuden päätietokannan. Kuormitusta tasataan horisontaalisesti hajauttamalla tieto järjestelmän eri lohkoille, joista jokainen sisältää yhden päätietokannan ja yhden tai useamman alitietokannan. Palvelimia, joilla lohkot sijaitsevat, voi olla järjestelmässä useita. Tällä tavoin ja tiedon kahdentamisella eri lohkoihin voidaan estää palvelukatkos laiterikkotilanteessa. Myös MongoDB käyttää MapReduce-tekniikkaa tehtävien suorittamiseen ja yhdistämistoimenpiteisiin. [12.]

4.1.3 Virtausprosessointi

Virtausprosessointi perustuu SIMD-malliin, jossa yhtä komentoa suoritetaan samanaikaisesti usealle eri tietolähteelle [13]. Virtausprosessoinnissa analysointi on esitellyistä malleista kaikista nopeinta, sillä se tapahtuu reaaliaikaisesti. Äärimmäisen nopeaa analysointia tarvitaan esimerkiksi osakekaupankäynnissä tai sotilaallisten toimenpiteiden hallinnassa. Myös mobiilisovellusten ja internet-pelaamisen myötä tarve reaaliaikaiselle vasteelle on kasvanut huomasti. Virtausprosessointimallia käyttävä järjestelmä voidaan usein yhdistää esimerkiksi rinnakkaisprosessointijärjestelmään tai tietokantaan. Tällöin tieto tallennetaan tiedostojärjestelmään, mutta välitöntä reagointia vaativan tiedon prosessointi suoritetaan ennen tallennusta virtausprosessijärjestelmässä. Esimerkiksi Twitterin kehittämä avoimen lähdekoodin Storm-järjestelmä hyödyntää reaaliaikaista virtausprosessointia.

Storm-klusteri on samankaltainen kuin Hadoop-klusteri, mutta Hadoopissa ajetaan MapReduce-tehtäviä ja Storm-järjestelmässä käytetään topologia-ajaja. Näiden erona on se, että MapReduce-tehtävä päättyy joskus, mutta topologiaprosessi prosessoi tietoa jatkuvasti. Storm-klusterin arkkitehtuurissa on päänoodeja ja alinoodeja. Päänoodilla on Nimbus-niminen taustaprosessi, joka jakaa koodia ympäri klusteria, nimeää tehtäviä laitteille ja valvoo virhetilanteita. Alinoodeilla on taustaprosessi nimeltä Supervisor, joka ottaa vastaan Nimbus-prosessin alinoodille nimeämät tehtävät. Näiden prosessien välistä koordinoitua tehdään niiden väliin asennettavalla Zookeeper-klusterilla, joka huolehtii myös prosessien toiminnasta. Supervisor ohjaa annettujen tehtävien mukaisesti noodin prosesseja, jotka suorittavat klusterilla käynnissä olevien topologia-ajajien osia. [14.]

Storm-käsittelyn läpi virtaava informaatio voidaan suoraan muuttaa haluttuun muotoon, esimerkiksi sarjasta twiittejä voidaan tuottaa reaaliaikaista статистиikkaa niiden aiheista. Näiden muutosten taustalla on spout- ja bolt-prosessit, joista spout-prosessi toimii tietovirran lähteenä virtausprosessille. Bolt-prosessi osakäsittelee sisään tulevan tietovirran ja jakaa sen pienempiin virtausosiin. Monimutkaiset muutokset vaativat useamman tason bolt-prosessointia. Spout- ja bolt-prosessien verkostot muodostavat topologia-prosessin, joka sisältää sen suorittaman analyysin raamit (kuva 5). [14.]



Kuva 5. Topologiaprosessikaavio [14].

4.2 Tallennusympäristö

Big data -tallennusympäristön rakentamisessa on useita eri vaihtoehtoja niin arkkitehtuurillisesti kuin tuotteellisesti. Tietyt lähtökohdat on kuitenkin hyvä muistaa kokonaisuutta suunnitellessa. Tallennusympäristön on oltava riittävän mukautuva sekä kapasiteetiltaan että nopeudeltaan. Koska yrityksen on erittäin vaikeaa ennakoida kasvavaa kapasiteetin tarvetta pitkälle tulevaisuuteen, järjestelmän kyky kasvaa joustavasti tarpeen mukaan ilman huomattavia kustannuksia tai suorituskykyrajoitteita, esimerkiksi tietoliikenteessä, on tärkeässä roolissa. Tallennustilan tulee olla jaettu tasoihin, jotta informaatio voidaan tallentaa järjestelmään hierarkkisesti siten, että eniten käytettävät tietokokonaisuudet ovat helpoiten saatavilla ja käyttävät nopeaa tallennustilaa. Harvemmin tarvittava tai vanha, pitkäaikaisesti säilytettävä informaatio tulee taas voida

tarvittaessa arkistoida. Hyvässä big data -tallennusympäristössä tieto voi siirtyä kerrokselta toiselle tarpeiden ja vaatimusten mukaisesti. [15.]

Hankittu tallennusympäristö on yleensä usean sovelluksen ja käyttäjän käytössä samanaikaisesti. Toiset sovellukset kykenevät automaattisesti siirtelemään tietoa kerrokselta toiselle, mutta kaikkiin ei ole ohjelmoitu vastaavaa toiminnallisuutta. On tärkeää, että tallennusympäristöllä on valmiudet hallinnoida ja kyky siirrellä sillä sijaitsevaa tietoa tallennuskerroksien välillä. Näin tallennettu sisältö on helposti ja laajasti saatavilla suuresta koosta ja pitkästä säilytysajasta huolimatta. Tallennusympäristön on hyvä tukea analyttisiä ja sisällöllisiä sovelluksia, sekä muodollista ja epämuodollista sisältöä. Lisäksi sen on tuettava automaattista työnkulku eli analytiikan siirtoa sovelluksesta sovellukseen tai käyttäjältä toiselle. Uuden tallennusympäristön on oltava joustava ja integroitavissa yrityksen olemassa oleviin sovelluksiin ilman merkittäviä kehitys- tai päivitystarpeita. Integroitavuuden on käsitettävä myös julkiset, yksityiset sekä hybridit pilviympäristöt. Oleellista on myös se, että tallennusympäristö kykenee itsenäisesti selviytymään eritasoisista vikatilanteista. Jos ympäristön joku resurssi vikaantuu, keskeneräiset työt tulee ohjata automaattisesti toiselle resurssille ilman merkittävää käyttökatoa. [15.]

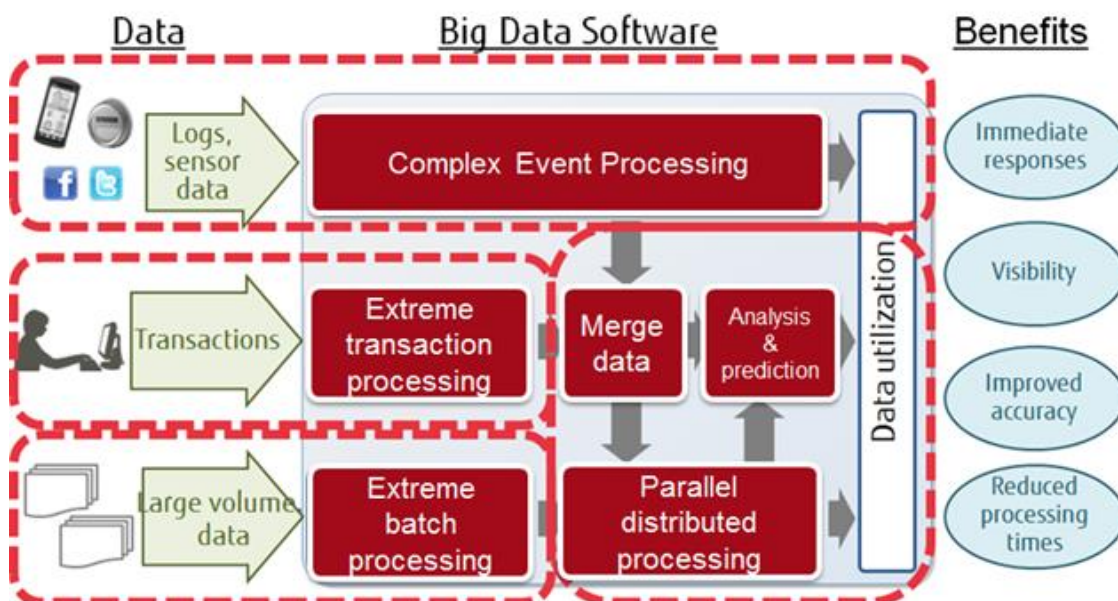
4.3 Pilviteknologia

Pilviteknologiassa asiakkaan teknisten ratkaisujen kehittäminen ja käyttö tapahtuu siten, etteivät palvelun käyttäjät voi nähdä tai hallita sen teknisiä yksityiskohtia. Pilviteknologiassa palveluita ja resursseja tuotetaan, käytetään ja toimitetaan yrityksen ulkopuolisista verkoista, jolloin niistä samalla poistuu alusta- ja laitetyyppisidonnaisuudet. Pilvessä tietotekniset palvelut ovat hajautettuja palveluntarjoajan palvelimille ja niistä maksetaan käytön mukaan. Palvelut ovat heti saatavilla ja laajennettavissa tarpeen mukaan. Pilviteknologiassa asiakkaalle tarjotaan enemmän tehoa pienemmällä vaivalla ja ilman perinteisiä perustamiskustannuksia. Erilaisia pilvityyppejä on useita. Julkinen pilvi sijaitsee muiden palveluntarjoajan asiakkaiden kanssa jaetuilla palvelimilla, yksityisessä pilvessä infrastruktuuri muodostaa suljetun verkon, joka on varattu vain kyseisen pilven käyttöön. Hybridipilvi on sekoitus edellisistä. Koska infrastruktuuri sijaitsee palveluntarjoajan tiloissa, eikä asiakkaalla ole siihen näkymää, tietoturvallisuus on täysin palveluntarjoajan varassa. Tietoa ei välttämättä pystytä suojaamaan kansallisilla tai EU-tasoisilla säädöksillä. [16.]

Pilviteknologia toimii alustana monille big data -palveluille ja ympäristöille. Harvalla yrityksellä on olemassa olevaa infrastruktuuria tai valmiuksia investoida kerralla big datan edellyttämään tallennustilakapasiteettiin ja laitteistoon. Osalla pilvipalveluiden tarjoajista on jo Hadoop-tuetut ympäristöt, mutta tuen asentaminenkaan ei yleensä ole prosessina monimutkainen. Lisäksi tietoa siirtyy jatkuvasti enemmän pilveen pilvipalveluiden yleistyttyä. Tästä johtuen tiedon analysointi on luonnollista suorittaa siellä missä tieto sijaitsee. [17.]

5 Tiedonkeruu ja avustavia teknologioita

Tieto kerätään erilaisilla menetelmillä joko suoraan tai välillisesti sen lähteestä ja vietään big data -ympäristöön tallennettavaksi ja käsiteltäväksi, jotta tuloksena saadaan lisäarvoa tuottavia ennusteita ja raportteja. Perinteisesti tieto kerätään palvelinlaitteiden kirjoittamista lokitiedostoista tai muista vastaavista tekstitiedostoista keskitetysti, jolloin tieto on usein rakenteellista ja sen käsittely on johdonmukaista. Tiedon keräys voi tapahtua usealla eri menetelmällä tai teknologialla. Monilla lokienhallintajärjestelmillä on omat agenttiohjelmistonsa, jotka asennetaan kohdelaitteelle välittämään lokitietoa. Ne toimittavat tiedon emolaitteelle käsiteltäväksi, mieluiten salattuna. Tiedon kerääminen laitteiden lokikannasta on perinteisin tiedonkeruun muoto tietoteknisissä ympäristöissä, mutta tietoa voidaan kerätä muillakin tavoin lähes mistä tahansa. Esimerkiksi antureiden sekä muiden mittarien käyttö mahdollistaa olosuhdemuutosten mittauksen ja muuttamisen digitaaliseen muotoon. Manuaalistakaan työtä ei sovi unohtaa, vaikka nykyään paljon onkin automatisoitu. Monet tietolähteet ja järjestelmät vaativat edelleen tiedon käsityötä. Kuvassa 6 on esitetty tietoyksikköjen kulku erilaisten prosessointimenetelmien läpi ja sen käsittelyn komponentteja big data -ympäristössä. Tiedonkeruuseen liittyviä menetelmiä ja ratkaisuja on esitelty seuraavaksi.



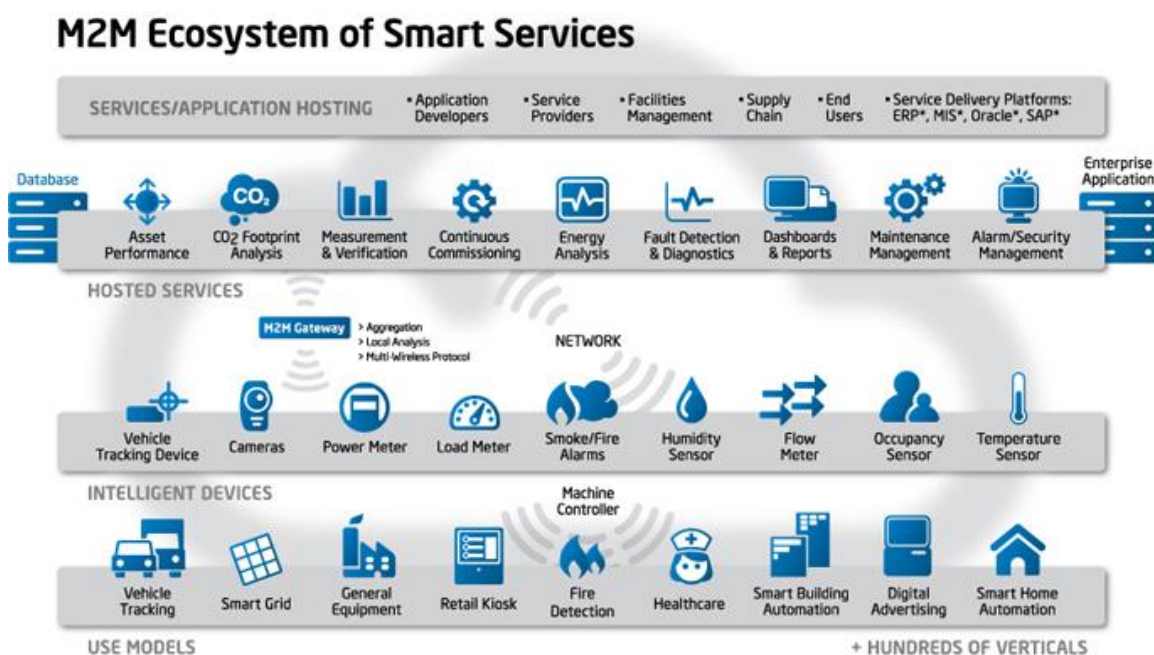
Kuva 6. Big data -konseptin käsitteellinen arkkitehtuuri [8].

5.1 Kaupallinen tieto

Organisaatiot käyttävät usein oman sisäisen informaation lisäksi myös muiden organisaatioiden, kuten julkishallinnon, tutkimuslaitosten tai sosiaalisen median keräämää tietoa. Muiden organisaatioiden tuottama raakatieto saattaa olla hankalaa ja kallista kerätä tai käsitellä, mikä on avannut markkinat tiedon kaupankäynnille. Useat yritykset tarjoavat valmiiksi esikäsiteltyä tietoa useista eri tietolähteistä korvausta vastaan. Vielä toistaiseksi tietoa markkinoivat yritykset ovat kehittyviä, eikä selkeää liiketoiminnallista suuntaa ole. Pelkän tiedon myynti ei aina ole kaupallisesti kannattavaa, vaan yritysten tulisi laajentaa toimintaansa tarjoamaan myös tiedonkäsittelyn palveluita. Monet tietopakettien tarjoajat ovat jo huomanneet haasteet siinä, kuinka tehdä liiketoiminnasta kannattavaa ja miten kääntää kasvava tiedon määrä tuottoisaksi liiketoiminnaksi. Toinen haaste löytyy verkon toimintanopeudesta. Tiedon käsittely ja etsiminen tulisi saattaa kestoltaan tasolle, joka tuo lisäarvoa tietoa hyödyntävän yrityksen liiketoiminnalle ilman merkittäviä tietoliikenteellisiä lisäkustannuksia. Lähitulevaisuudessa tulemme näkemään trendin, jossa tiedontarjoajat siirtyvät teknologisesti noudattamaan big data -filosofiaa ja pilviteknologian palveluntarjoajat laajentavat liiketoimintansa koskemaan myös tietovarastotarjontaa lisäarvoa tuottavana palveluna. [18, s. 11–12.]

5.2 Laitteiden välinen langaton koneviestintä

Perinteisesti lokitiedon on ajateltu tulevan järjestelmälokien lisäksi esimerkiksi luottokorttihankinnoista tai internetissä hakukoneeseen syötetyistä hakusanoista, linkkien painamisista ja selaimen sivuhistoriasta. Yhä suurempi osa big data -analyysissä käytettävästä tiedosta tulee kuitenkin yksittäisistä, itsenäisistä laitteista, jotka keräävät tietoa ja välittävät sen toiselle laitteelle koostamista, analysointia ja säilytystä varten. Tästä kokonaisuudesta käytetään termiä M2M (machine-to-machine). M2M:n tuottama informaatio on todella hajanaista, ja sen määrä kasvaa valtavan nopeasti. Lisäksi tämän informaatiomassan mahdollinen lisäarvo on ymmärretty vasta hiljattain, mistä johtuen yritykset ovat alkaneet käydä uudestaan läpi jo kerättyä M2M-tietoa oppiakseen hyödyntämään tätä sektoria tiedon lähteenä entistä paremmin. [18, s. 16–17.]

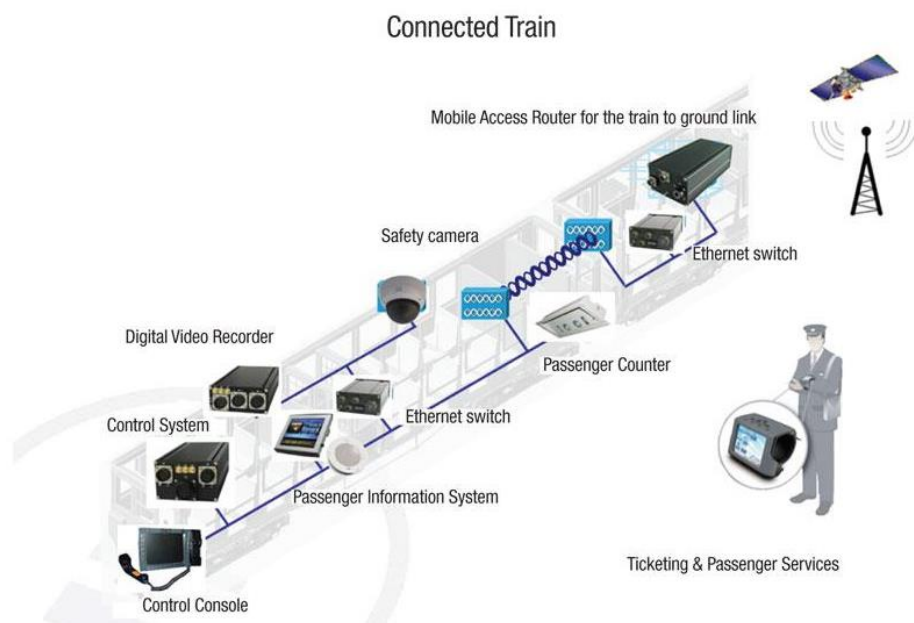


Kuva 7. M2M-ekosysteemi [19].

Kuvassa 7 on kuvattu älykkäiden laitteiden M2M-ekosysteemi. On arvioitu, että vuonna 2014 maailmassa on 150 miljoonaa M2M-yhteyttä ja 50 miljardia yhteyttä vuoteen 2020 mennessä [19]. Esimerkiksi tietullijärjestelmät perustuvat usein siihen, että ajoneuvon tuulilasiin on kiinnitetty RFID-tunniste, joka on liitetty luottokorttiin tai pankkitiliin ja koko maksuprosessi on automatisoitu. Ajoneuvon RFID-tunniste siirtää tiedot tietokantapal-

velimille, jotka tarkistavat tilin statuksen. Tietokantapalvelimet antavat luvan tietullimak-
sulle ja siirtävät maksun ajoneuvon haltijan tililtä. Järjestelmä tallentaa myös ajankoh-
dan, ajoneuvon tiedot ja omistajan identiteetin tietokantaan. Tietullipisteiden keräämää
informaatiota voidaan käyttää mm. ruuhka-aikojen tunnistamiseen, hienostuneempien
ja tarkempien maksujärjestelmien kehittämiseen. Näiden tietojen avulla taas voidaan
ennalta ehkäistä ruuhkahuippuja, jolla vastaavasti on epäsuora vaikutus polttoaineen
kulutukseen sekä saasteiden määrään. [18, s. 18–19.]

Tarkkailemalla materiaalin liikkumista siihen liitettyjen RFID-tunnisteiden avulla, voi-
daan selvittää muun muassa työntekijöiden tuottavuusastetta, varastovirheiden syitä ja
työnkulussa toistuvia kaavoja. M2M toimii apuvälineenä, kun päivittäisestä perustoi-
minnasta halutaan syvempää ymmärrystä. Sitä käytetään esimerkiksi Yhdysvaltojen
postipalvelussa lähetyksien lajittelussa ja lentoyhtiöissä matkalaukkujen automaatti-
seen siirtämiseen lentokentällä. Kuluttajat taas ovat alkaneet käyttää esimerkiksi mobiili-
sovelluksia kotivalojen kytkemiseksi päälle etänä tai televisio-ohjelmien tallentamiseksi.
Vähittäiskaupassa voidaan hyödyntää älykkäitä hyllyjärjestelmiä, jotka tunnistavat,
kuinka pitkään asiakas viipyy tietyn tuotteen kohdalla. Kerätyn tiedon perusteella voi-
daan tutkia, minkälainen esillepano saa asiakkaan tekemään ostopäätöksen. [18, s.
18–19.]



Kuva 8. M2M-teknologia raideliikenteessä [20].

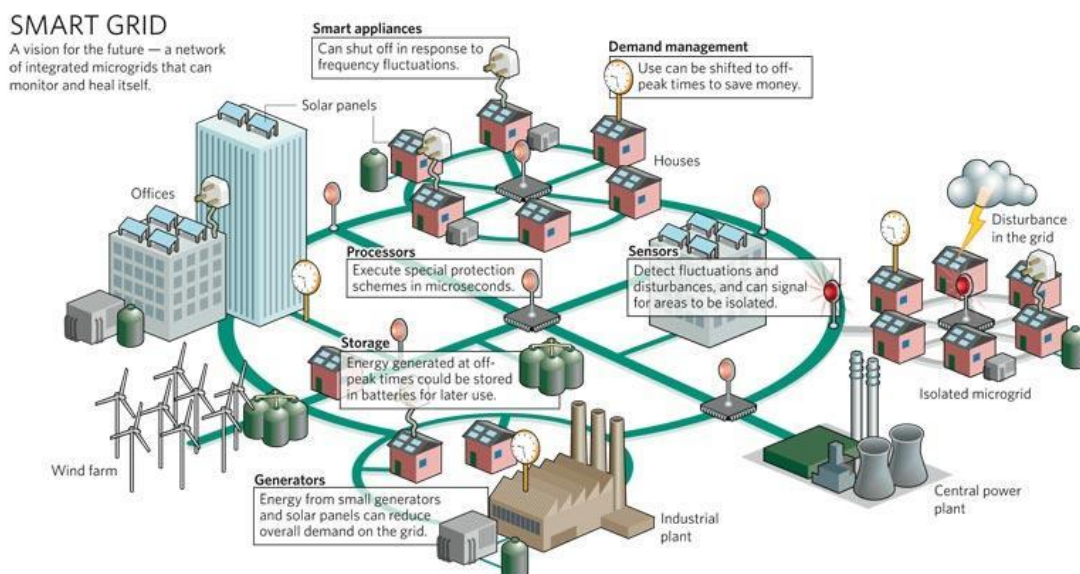
Kuvassa 8 on kuvattu M2M-teknologian käyttöä raideliikenteessä. Tavoitteena on kerätä riittävästi tietoa, jota analysoimalla matkustajien siirtyminen kohteiden välillä sujuisi mahdollisimman miellyttävästi. M2M:n avulla voidaan mm. päivittää reaaliaikaisesti liikennevälineiden lähtö- ja saapumisaikoja, jolloin myös jatkoyhteyksien hallinta tehostuu ja on käyttäjäystävällisempää. [20.]

M2M tuottaa valtavan määrän tietoa, jonka käsittely ja tallennus ovat suurimpia haasteita big data -konseptissa. NykYTEknologialla kaikkea kerättyä tietoa ei ehditä käsittelemään ja tallennustilan kasvava tarve on sekä teknisesti että taloudellisesti haastavaa. M2M tulee olemaan avain toimintojen yksityiskohtaiseen ymmärtämiseen liiketoiminnan näkökulmasta. Tämän seurauksena se saattaa joutua tietoturvanäkökulmasta huonoon asemaan. Sovellukset, joita käytetään esimerkiksi autojen käynnistykseen ja rakennuksien sisäänpääsyyn, ovat tietorikollisen näkökulmasta houkuttelevia kohteita. Tiedonsiirto edellyttää salausmenetelmää, joka ei kuormita tiedonsiirtokaistaa, ja laitteiden tietoturvalliset tunnistusmenetelmät saattavat vaikeuttaa laitteiden liittämistä toisiinsa. Suurin haaste tulee kuitenkin olemaan M2M-laitteiden turvallisuuden ylläpito, sillä laitteita on todella suuri määrä. Tietoturva- tai laitepäivitykset tulevat olemaan ongelmallisia toteuttaa vaaditussa ajassa, mutta myös valvonta sormeilun tai murtautumisen varalta voi olla hankalaa. [18, s. 23–24.]

5.3 Älykäs sähköverkko

Huoli fossiilipohjaisten polttoaineiden riittävyydestä ja niiden käytön ympäristöhaitoista on johtanut uusiutuvien energiamuotojen kasvavaan kysyntään. Tuuli- ja aurinkovoiman tuotto vaihtelee kuitenkin niin paljon, että sähkönjakelun ja tuotannon hallinnalle on tarvittu hienostuneempia ohjaussysteemejä. Älykäs sähköverkko eli smart grid käyttää tieto- ja kommunikointiteknologiaa kerätäkseen tietoa toimittajien ja kuluttajien käyttäytymismalleista. Kerättyä tietoa analysoimalla voidaan parantaa tehokkuutta, luotettavuutta, taloudellisuutta sekä sähköön tuotantoa ja jakelua. Smart gridissä digitaalinen teknologia on yhdistetty sähköverkkojen kanssa tiedon keräämiseksi erilaisten mittareiden avulla. Tiedon perusteella voidaan tutkia ja optimoida sähkönkäyttöön liittyviä asioita. Laitteistot, kuten ilmastointi, jäähdyttimet tai lämmittimet, voidaan säätää välttämään sähkönkulutuksen huippuaikoja tai keskittymään ajankohtiin, jolloin uusiutuvaa energiaa on saatavilla. Lisäksi sähkölaitteita voidaan priorisoida, jolloin alhaisen prioriteetin laitteet käyttävät energiaa vain sen ollessa halvimmillaan. Kuvassa 9 on älyk-

kään sähköverkon tulevaisuuden visio, jossa laitteet kommunikoivat keskenään ja toiminnot ovat vikatilanteissa itsestään korjautuvia.



Kuva 9. Älykäs sähköverkko [21].

Smart grid -tekniikan muodostavat kehittyneet mikroprosessorimittarit eli smart meterit, mittareiden lukuvarusteet, valvontaratkaisut, analyysityökalut ja kehittyneet kytkimet sekä kaapelit. Laitteiden välinen kommunikointi on reaaliaikaista, jolloin järjestelmän luotettavuus ja verkon kuormitus voidaan pitää optimaalisena. Sähköverkon stabiiliutta arvioidaan jatkuvasti ja laitteiston käyttöaste pystytään pitämään korkeana valvonnan kautta. Näillä menetelmillä voidaan myös ehkäistä energiavarkauksia ja tukea ohjausstrategioita. Suurimmat haasteet kohdistuvat smart meter -tekniikkaan ja sosiaalisiin näkökulmiin. Kuluttajien huolenaiheena on se, voidaanko sähkökäyttötietoa käyttää heitä vastaan tai kärsiikö sähköjakelun oikeudenmukaisuus käytön korkeasta yksilöintiasteesta. Lisäksi hinnoittelumekanismi saattaa olla epäselvä ja tekniikka antaa hallitukselle välineet kontrolloida kaikkia sähkökäyttäjien toimintoja. [22.]

5.4 Mittaustietojen hallintajärjestelmät

Älykkäiden mittauslaitteiden yleistytessä haasteeksi on muodostunut se, miten kaikkea kerättyä informaatiota voidaan käsitellä ja hallita ilman, että niitä varten tarvitsee hankkia arvokkaita mittarinlukijoita. Mittareiden tuottama informaatio ei ole pelkästään ener-

giankäyttöön liittyvää, vaan se sisältää myös muihin suureisiin, kuten jännitteeseen tai lämpötilaan liittyvää tietoa. [18, s. 53.]

Mittaustietojen hallintajärjestelmä MDMS eli Meter Data Management System on älykkäiden sähköverkkojen avainkomponentti, jonka avulla tiedot kerätään älykkäiden mittarien keräysjärjestelmistä esikäsittelyä ja tallennusta varten. Tiedot validoidaan ja siirritään järjestelmässä ennen kuin ne siirtyvät laskutus- ja analyysijärjestelmien käyttöön. MDMS tuottaa informaatiota, jonka avulla voidaan tutkia kuormitusta ja kysyntää sekä tuottaa asiakaspalvelun käyttöön metristä tietoa. Se tarjoaa rajapintoja sovellusten ohjelmoinnille, jotta mittareista kerätty tieto voidaan prosessoida myös muiden resurssien käyttöön. Ekologisen teknologian tutkimusjaosto GTM Research ennustaa, että MDMS-markkinat tulevat kasvamaan yli 300 prosenttia vuonna 2014. [23.]

6 Tiedon ja sen analysoinnin vaatimukset

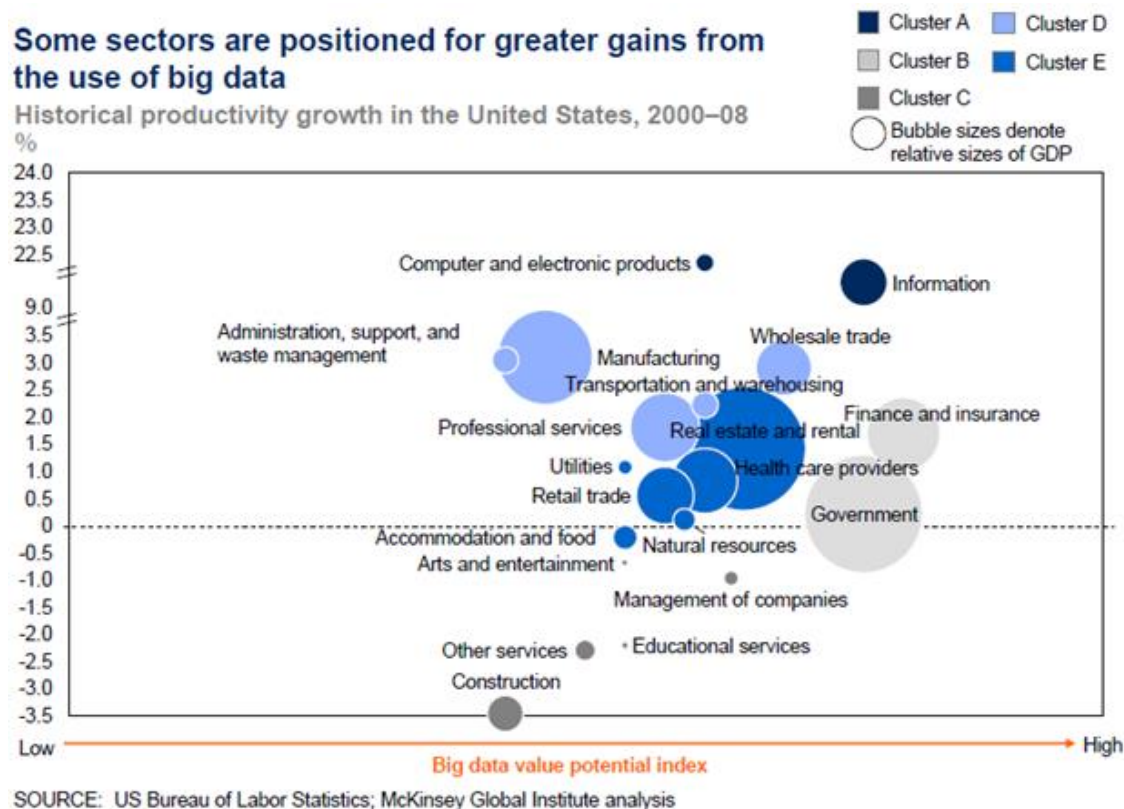
Tiedon analysoinnissa tärkeimpänä lähtökohtana on selvitys siitä, mitä halutaan tutkia ja minkälaisessa muodossa tulokset kannattaa tuottaa. Esimerkiksi tutkittaessa tuotteen myynnin vaihteluita sesonkien mukaan järkevää olisi tuottaa graafinen raportti, kun taas bakteerien määrä eri tuote-erissä lienee selkeintä raportoida taulukkomuotoisena. Tämän jälkeen voidaan määritellä ne komponentit, joiden seuraaminen tutkimuksessa on analyysin kannalta merkittävää. Samalla on selvittävää, millä tasolla analyysi halutaan tehdä. Kannattaako tulokset esittää lähinnä tilastollisesta näkökulmasta vai tulisiko analysoinnissa käyttää kvalitatiivisempia menetelmiä, jotta tulokset vastaavat selvityksen vaatimuksia riittävän tarkasti? Analyysissä tutkitaan systemaattisesti tallennettua tietoa siltä osin kuin tutkimuskohteen ja analyysin tason kannalta on oleellista. Tapahtumia ketjutetaan, yhdistetään tai ryhmitellään siten, että niiden perusteella voidaan tehdä asiaankuuluvia johtopäätöksiä. Jotta tiedon analyysi tuottaa odotettua lisäarvoa, on huomioitava tiedon laadukkuus, eheys ja riittävyys. Tiedon eheys on merkittävää etenkin, kun kyse on tietoturvaratkaisuista tai todistustaakasta. Tietopohjan muuttuminen saattaa kuitenkin vaikuttaa ylimalkaisemmankin analyysin tuloksiin, jos virheellistä tietoa on prosentuaalisesti merkittävä määrä. Tästä syystä on hyvä pyrkiä säilyttämään tiedon muuttumattomuus läpi koko tiedon tallennuksen ja käsittelyn elinkaaren.

Kerättävän tiedon kriittisin vaatimus on sen laadukkuus. Tiedon laadun mittareita ovat muun muassa tarkkuus, kokonaisuus, yhdisteltävyys, tuoreus ja ainutlaatuisuus. Laadukas tietoyksikkö pitää sisällään kaiken analysoinnin kannalta oleellisen informaation ja siitä on karsittu tarpeettomat osat pois. Se on yhdistettävissä muihin tietoyksiköihin, tietosisältö ei ole vanhentunutta ja se esiintyy tietojärjestelmässä vain yhden kerran. Jo ennen big data -konseptin ilmestymistä tiedon laatu on ollut merkittävässä roolissa yrityksen liiketoiminnan kannalta. Turhan tai virheellisen tiedon siivoaminen on ollut sekä tallennustilaa säästävää että tietojärjestelmien käyttöä selkeyttävää. Tiedon laadukkuus on myös big data -konseptissa tärkeää, vaikka laadukkuuden takaaminen onkin haasteellista laajan tietopohjan vuoksi. Samaa tietopohjaa voidaan käyttää vastamaan useaan eri kysymykseen ympäri organisaatiota tai jopa organisaatorajat rikkoen, jos se on huolellisesti jäsenneltyä ja laadukasta. Big data -analysoinnin ja tallennustilan alentuneiden kustannusten myötä tietoa on alettu kerätä perinteisten lokilähteiden lisäksi myös esimerkiksi mobiililaitteista ja erilaisista sensoreista, mikä on kasvattanut laatupaineita entisestään. Tiedon harkitsematon kerääntyminen eli obesiteetti saattaa olla yritykselle kohtalokasta ilman kunnollista karsintastrategiaa. Mitä enemmän lähesytään big dataa konseptina ja sen täyttä hyödyntämistä, sitä enemmän tallennettavaa tietoa on. Vaikka tallennustilakustannukset laskevatkin, tiedon käsittely, siistiminen ja analysointi saattavat tuoda merkittäviä lisäkustannuksia jos niitä ei tehdä suunnitelmallisesti. Etenkin tiedon obesiteetista ja huonolaatuisuudesta yhdessä voi organisaatiolle muodostua pysyvää haittaa. Paras tapa välttää näitä onkin hyödyntää alusta saakka tietoalueen parhaita käytäntöjä ja suunnitella strategia niiden käyttämiseksi. [18, s. 7–10, s. 43.]

Riittävän tekniset ja analyttiset taidot hallitsevat henkilöt voivat analysoida tietopohjaa suoraankin, mutta markkinoilla on useita järjestelmiä, jotka on suunniteltu tuottamaan analyysijä automaattisesti. Automaattiset järjestelmät tulevat kyseeseen etenkin silloin, kun tietopohja on laaja tai tuloksia tarvitaan nopeasti. Analyysijärjestelmän valinnassa pätevät samat lainalaisuudet kuin tiedon käsittelyarkkitehtuurin rakentamisessakin. Järjestelmän on oltava käyttötarkoitukseensa sopiva ja integroitavissa olemassa oleviin ympäristöihin ja järjestelmiin. Sen kustannuksien, vaatimusten ja ominaisuuksien tulee olla tasapainossa yrityksen tarpeiden ja resurssien kanssa. Tärkeintä on, että hankittava järjestelmä vastaa toiminnan asettamia vaatimuksia mahdollisimman kattavasti sellaisenaan, jotta vältetään ylimääräiset järjestelmän integroimisesta ja räätälöinnistä aiheutuvat kustannukset sekä voidaan minimoida käsin tehtävän työn määrä.

7 Sovellusalueita ja esimerkkejä

Big datan hyödyntämisen potentiaali vaihtelee toimialoittain merkittävästi. Kuvassa 10 on kuplakaavio Yhdysvaltojen eri sektoreiden mahdollisuuksista hyödyntää big dataa toiminnan tehostamiseen.



Kuva 10. Big datan hyödyntämispotentiaali toimialoittain [1, s. 10].

Vaikka tutkimus on toteutettu Yhdysvalloissa, linjaus on vastaava koko länsimaisessa yhteiskunnassa. Toimialueen suhteellista osuutta bruttokansantuotteesta kuvataan kaaviossa kuplan koolla ja värillä liitääntä toimialaklusteriin. Toimialat on jaettu klustereihin siten, että A-klusteriin kuuluu tieto- ja elektroniikka-alat, B-klusteriin rahoitus- ja vakuutusala sekä hallinto, C-klusteriin ne alat, joissa tuottavuus on laskenut merkittävästi, D-klusteriin alat, joissa tuottavuus on noussut ja E-klusteriin pääsääntöisesti paikalliset palvelut. Big datan käytön potentiaali-indeksi toimialoilla on laskettu käyttäen apuna viittä mittaria: analysoitavan tiedon määrää, suorituskyvyn vaihtelua, sidosryhmien keskimääräistä lukumäärää, liiketoiminnan intensiteettiä ja sektorin luontaista epävakautta. [1, s. 9–10.]

Kuvasta 10 voidaan havaita, että suurimmat hyödyntämismahdollisuudet big datalle löytyvät B-klusterista eli hallinnosta sekä vakuutus- ja rahoitusosalta. Merkittävää potentiaalia on havaittavissa myös kiinteistöalalla, terveydenhuollon sektorilla sekä vähittäis- ja tukkumyynnissä. Potentiaalin painottuminen edellä mainittuihin ei kuitenkaan tarkoita, ettei big data -analytiikasta olisi muilla aloilla hyötyä, sillä mittareina on käytetty myös tämänhetkistä valmiutta analytiikan käyttöön. Seuraavana on esitelty eräiden sovellusalueiden mahdollisuuksia ja tapoja hyödyntää tiedon analysointia.

7.1 Liiketoiminnan tehostaminen

Operational intelligence eli OI on liiketoiminnan analytiikkaa, joka tuo läpinäkyvyyttä ja syvällisempää ymmärrystä tietoon, tapahtumavirtaan ja liiketoiminnan toimintaan. OI-ratkaisuissa tehdään kyselyitä tietovirtaa ja tapahtumatietoa vastaan, jotta saadaan tuotettua reaaliaikaisia tuloksia välitöntä toiminnanohjausta varten. OI:n tarkoitus on monitoroida liiketoiminnan tapahtumia ja tunnistaa tilanteet, joissa ilmenee tehottomuutta, mahdollisuuksia tai uhkia. Tilanteiden tunnistamisen jälkeen OI-järjestelmä tarjoaa ratkaisuja tilanteen hallitsemiseksi. OI:n avulla voidaan tehostaa liiketoimintaa, tutkia miten odottamattomat tapahtumat, esimerkiksi IT-ympäristössä, vaikuttavat siihen ja kuinka liiketoiminnalliset tapahtumat tukevat liikevaihdon kasvua tai pienenemistä. Tämä saavutetaan tarkkailemalla toiminnan edistymistä, tutkimalla useita mittareita ja välittämällä tietoa eteenpäin reaaliajassa. Lisäksi nämä mittarit toimivat aloituspisteinä jatkoanalyysille, jossa voidaan porautua syvemmälle järjestelmään juurisyyyn paikantamiseksi sitomalla poikkeavuudet tiettyihin toimenpiteisiin ja liiketoiminnallisiin tapahtumiin. [24.]

Erilaiset OI-ratkaisut saattavat käyttää eri teknologioita ja olla ympäristöön implementoituja eri tavoin. Tavallisia OI-ratkaisuja ovat esimerkiksi reaaliaikainen monitorointi tai tapahtumien perusteella tehtävät korjausliikkeet, mutta myös big data -analytiikka. OI-ratkaisuilla voidaan monitoroida ja analysoida monimutkaista ja laajaa tietopohjaa. OI-järjestelmiä verrataan usein BI-järjestelmiin, sillä molemmat tuottavat liiketoiminnallista analyysiä suuresta tietomäärästä. OI-järjestelmät keskittyvät ensisijaisesti toiminnan tuottamiin seurauksiin, BI-järjestelmät sen sijaan keskittyvät ensisijaisesti historialliseen tietosisältöön. OI on reaaliaikainen järjestelmä kun taas BI-järjestelmissä usein tarvitaan heräte, jotta kaavoja voidaan tunnistaa. Markkinoille on tullut myös reaaliaikaisia BI-järjestelmiä, mutta niiden toiminta perustuu perinteisissä relaatiotietokannoissa ole-

vaan tietosisältöön. OI ja BI eivät kuitenkaan ole toisiaan poissulkevia järjestelmiä, OI tuo menetelmiä nopeaan päätöksentekoon, kun taas BI-järjestelmät sopivat pitkän ajan suunnitteluun. [24.]

7.2 Markkinointi

Big datan hyödyntäminen vähittäiskaupassa tuo uusia mahdollisuuksia markkinointiin ja etenkin kohdennettua markkinointia voidaan kehittää sen avulla huomattavasti. Tähän saakka kohdennetulla markkinoinnilla on tarkoitettu lähinnä iän, sukupuolen tai asuinpaikan perusteella tehtävää markkinointia, mutta tarkemman analysoinnin kautta kohdennus voidaan viedä vieläkin yksilöllisemmäksi. Joka kerta kun henkilö tekee ostoksia, hän jättää jäljen ostostottumuksistaan kaupan tietojärjestelmään. Näitä tietoja analysoimalla kaupat pystyvät seuraamaan sekä yleisiä markkinatrendejä, että yksittäisen henkilön kulutustottumuksia hyvinkin tarkasti. Etenkin muutokset kulutustottumuksissa merkitsevät usein muutoksia muilla elämänalueilla. Näihin muutoksohtiin markkinastrategioissa pyritään reagoimaan ensimmäisenä, jotta kulutustottumuksia saadaan ohjattua liiketoiminnalle suotuisaan suuntaan.

Amerikkalainen vähittäismyyjä Target alkoi vuonna 2002 selvittää algoritmia, jonka avulla voitaisiin tunnistaa raskaana olevat naiset. Tutkimukset olivat osoittaneet, että kuluttajan ostotottumuksia on äärimmäisen hankalaa muuttaa, mutta että jokaisen elämässä on vaihteita, jolloin vanhojen rutiinien asema heilahtelee. Yksi näistä hetkistä on tuore vanhemmuus, sillä vanhempien ollessa uupuneita ja tunnekuohun vallassa heidän ostosrutiininsa ja merkkiuskollisuutensa horjuu. Lapsen synnyttyä tiedosta tulee julkinen, jolloin monet yritykset yrittävät saada tuoreet vanhemmat osaksi asiakaskuntaansa erilaisin tarjouksin. Target halusi kehittää menetelmän, jolla raskauden todennäköisyys voitaisiin todeta huomattavasti ennen muita kauppiaita ja mielellään siten, että tulevaa äitiä voidaan lähestyä raskauden toisella kolmanneksella. Tutkimuksien mukaan tämä on ajankohta, jolloin tulevat äidit alkavat ostamaan uusia tuotteita, kuten ravintolisiä ja äitiysvaatteita. Jos tuleva äiti tunnistettaisiin oikeaan aikaan, voitaisiin oikeilla toimenpiteillä saada hänestä kanta-asiakas vuosikausiksi. Target oli jo vuosia kerännyt yksilöllisiä asiakastietoja jokaisesta kaupan kanta-asiakkaasta kanta-asiakasohjelmien, verkkovierailujen ja mielipidekyselyiden avulla. Tallennetuista tiedoista nostettiin esille ostohistoriatiedot kaikilta niiltä naisilta, jotka olivat ilmoittautuneet Targetin vauvarekisteriin. Tietoja analysoitiin, kunnes tuloksista oli havaittavissa käyt-

tökelpoisia kaavoja. Lopputilanteessa voitiin tunnistaa 25 tuotetta, joiden yhteisanalysoinnilla voitiin laskea todennäköisyysluku raskaudelle ja jopa melko tarkka arvio laskutulle ajalle. Algoritmin ansiosta todennäköisemmin raskaana olevia naisia voitiin lähestyä esimerkiksi lastenhoitovarustetarjouksilla ennen kilpailijoita. [25; 26.]

Usein verkkokaupat, kuten Amazon, hyödyntävät niin ikään big data -analytiikkaa markkinoinnissaan. Amazonilla käyttäjien toimintaa verkkokaupassa tutkitaan ja sen perusteella kullekin käyttäjälle suositellaan sopivia kirjoja yksilöllisesti [27]. Toinen suositelualgoritmia hyödyntävä yritys on videopalveluntarjoaja Netflix. Se on esimerkiksi varmistanut uusien sarjojensa menestystä tai jopa saanut ideoita uusille sarjoille analysoimalla sen 33 miljoonan käyttäjän videokäyttötymistä ja selvittämällä, mitä ihmiset haluavat katsoa. [28.]

7.3 Energiankulutus

Suurin osa sähköjakelun palveluntarjoajista Yhdysvalloissa on alkanut suunnitella tai toteuttaa strategiaa, jolla siirtyä digitaaliseen aikaan. Painopisteenä strategiassa on yhä enenevässä määrin laitteiden yhdistäminen älykkääseen sähköverkkoon, jotta voitaisiin vähentää sähkökatkoja, lisätä uusiutuvan energian käyttöä ja ylläpitää asiakkaiden mielenkiintoa sähköjakelijan toimintaa kohtaan. Laitteita asennetaan kaikille sähköverkon tasoille ala-asemista aina sähkön siirtoon ja jakeluun saakka. Monet aloittelevat yritykset pyrkivät markkinoimaan älykkäitä sähkönhallintajärjestelmiä kotitalouksillekin, mutta liiketoiminta ei vielä ole tehnyt läpimurtoaan. Kuluttajat eivät ole vielä riittävän kiinnostuneita oman kotinsa energiatehokkuudesta, että olisivat valmiita hankkimaan kohtalaisen kalliin monitorointijärjestelmän. Tilanteen ennakoita kuitenkin muuttuvan, mutta ajankohdasta on annettu vasta varovaisia arvioita. Rakennuksien valaistus-, lämmitys- ja jäähdytysjärjestelmien energiankäytössä on runsaasti parantamisen varaa. Lisäämällä laitteisiin älykkyyttä niitä voidaan verkottaa ja ohjata kommunikoidaan keskenään hallintajärjestelmien kautta. Esimerkiksi valaistuslaitteisiin asennetuilla langattomilla sensoreilla ja niistä saatavan tiedon hyödyntämisellä älykkäissä ohjausjärjestelmissä voitaisiin vähentää valaistuksen energiankulutusta jopa puolella. Tällaisia valaistuksen ohjausjärjestelmiä käytetään toistaiseksi vähän, vaikka valaistukseen käytettävän energian lisäksi vastaavalla järjestelmällä voitaisiin vähentää huomattavasti myös esimerkiksi lämmitys- ja jäähdytysjärjestelmien energiankulutusta. [29.]

Suomessa sähköverkkojen älykkyys on jo eräänlaisella perustasolla, vaikka laitteisto-kohtaiseen analytiikkaan ei ole vielä yleisesti edetty. Suomessa sähköverkosta voidaan esimerkiksi paikantaa viat automaattisesti. Lisäksi verkon käyttöä voidaan optimoida ja etäluettavia mittareita pyritään asentamaan laajalti. Vuoden 2013 lopussa vähintään 80 prosenttia jakeluverkkojen asiakkaista on tuntimittauksen ja etäluennan piirissä. Tämä helpottaa seuraamaan ja hallitsemaan omaa sähkönkäyttöä. Sähköyhtiö voi myös hyödyntää mittareita uusien palveluiden tuottamisessa ja asiakaspalvelun tehostamisessa. [30; 31.]

Elektronisiin laitteisiin, kuten ilmastointijärjestelmiin tai uuneihin voidaan liittää 3G-lähetin, joka siirtää keskitettyyn operointikeskukseen tietoa laitteen sähkönkulutuksesta. Operointikeskuksessa tieto analysoidaan, jotta voidaan selvittää sekä yksittäisten laitteiden kulutusta että henkilöiden kulutustottumuksia. Tällä tavoin voidaan tuottaa kodinomistajille tai kiinteistöjohtajille yksityiskohtaista tietoa energiankulutuksesta, jotta he voivat reagoida siihen kulutusta vähentämällä. Energiankulutukseen keskittyneen yrityksen GroundedPowerin toimitusjohtaja Paul Cole kehitti kotitalouksien käyttöön älykkään järjestelmän, jonka avulla pyrittiin vaikuttamaan energiankulutukseen analysoimalla sekä energiankulutusta että kuluttajakäytöstä. Lisäksi järjestelmän suunnittelussa hyödynnettiin psykologista näkökulmaa, sillä kilpailuasetelmaan joutuminen usein tehostaa yksilön pyrkimyksiä. Yritysjärjestelyjen jälkeen järjestelmän käyttö laajeni ja se asennettiin koeryhmän käyttöön vaikutuksien havaitsemiseksi. Järjestelmä perustui siihen, että asiakkaat pystyivät seuraamaan energiankulutustaan reaaliaikaisesti ja saamaan palveluntarjoajalta neuvoja kulutuksen vähentämiseksi. Järjestelmän kautta he pystyivät oman kulutuksensa lisäksi seuraamaan yleistä energiankulutuksen tasoa sekä samalla alueella asuvien, muiden koeryhmään kuuluvien kotitalouksien kulutusta. Näillä tavoin kotitaloudet saatiin vähentämään energiankulutustaan keskimäärin 7,8 prosenttia. [18, s. 20, s. 57–58.]

Perinteisten energialähteiden hupeneminen on kasvava ja globaalinen huolenaihe. Fossiilisten polttoainevarastojen käyminen vähiin on pakottanut kehittämään vaihtoehtoisia energiamuotoja, mutta varsinaisia läpimurtoja kehityksessä ei ole tapahtunut. Tuuli-, vesi- ja aurinkovoima sekä muut uusiutuvat energialähteet ovat jo käytössä, mutta niiden tuottama energia ei riitä kattamaan kysynnän kasvua. Tästä syystä energialähteiden kehittämisen rinnalla on pyrittävä myös hidastamaan energiatarpeen kasvua. Käyttämällä älykkäitä sähköverkoja, kulutusta voidaan hallita paremmin ja kohdentaa siten, ettei hukkaenergiaa pääse syntymään. Tässä kohtaa analytiikalla on tär-

keä rooli avaintekijöiden, kuten kysyntäpiikkien, jatkuvaa huoltoa vaativien tai tehottomien laitteiden tunnistamisessa, jotta energiakulutuksen pienentäminen olisi mahdollista.

7.4 Terveydenhuolto

McKinsey Global Institute arvioi, että Big data -palveluilla saavutettava potentiaalinen arvo terveydenhuoltosektorille voisi olla Yhdysvalloissa jopa 300 miljardia dollaria vuodessa. Taloudellisen hyödyn lisäksi analysoimalla terveydenhuollon toimintaa voitaisiin vähentää lääketieteellisiä erehdyksiä, ymmärtää potilasriskejä paremmin, tunnistaa tehokkaammin tilanne, jossa potilaalla on useampi yhdenaikainen tila ja huomata taudinaiheuttajien leviäminen aikaisemmassa vaiheessa. [1, s. 50.]

Lääkäri Jeffrey Brenner alkoi vuoden 2001 New Jersey'n Rutgers Universityn ampumistapauksen, jossa 21-vuotiaasta opiskelijaa ammuttiin nilkkaan epäonnistuneen huume-kaupan yhteydessä, jälkeen kiinnostua rikosinformaation analysoinnista poliisien toiminnan oltua tilanteessa epäammattimaista ja tehotonta. Hän alkoi tehdä omaa kartoitustaan alueen rikoksista keräämällä tietoa paikallissairaaloilta. Hän tuli havainneeksi, että ensiapupoliklinikoilla tarjottava hoito ei ollut asianmukaista eikä kustannustehokasta. Brennerin aloittamia tutkimuksia jatkettiin hänen perustamansa säätiön toimesta yli 600 000 poliklinikkakäynnin analysoimiseksi kahdeksan vuoden ajalta. Tuloksista havaittiin, että 13 prosenttia potilaista aiheuttivat 80 prosenttia kustannuksista ja tämä oli tullut maksamaan yhteiskunnalle yli 650 miljoonaa dollaria. Tuloksista voitiin myös päätellä, että suurin osa ongelmista olisi ollut ennakoitavissa oikeanlaisella toiminnalla ja turhien käyntien määrää olisi voitu vähentää. Ongelmakohdat paikantamalla pystyttiin rakentamaan järjestelmä, jonka avulla toimintaa saatiin tehostettua, käyntien määrää vähennettyä, hoidon laatua parannettua ja kustannuksia pienennettyä. [32.]

Missourin yliopistossa tehdyn tutkimuksen mukaan henkilön profiili ja toiminta sosiaalisessa mediassa saattavat antaa vihjeitä mielenterveysongelmista. Tutkimuksen mukaan profiiliin ja toiminnan analysointi voi antaa terapeutille kokonaisvaltaisemman kuvan henkilön terveydentilasta kuin perinteiset haastattelumenetelmät ja sitä voitaisiin jopa käyttää psykologien sekä terapeuttien työvälineenä. Sosiaalisen median toimintatavat antavat yksilöstä usein rehellisemmän ja luonnollisemman kuvan kuin nykyiset itsearviointimenetelmät. Vastaukset itsearvioinneissa ovat usein myös riippuvaisia hen-

kilön muistista, jonka tarkkuus saattaa vaihdella. Tutkimusta varten osallistujia pyydettiin tulostamaan toimintansa Facebookissa tutkimuksen käyttöön. Joillain henkilöillä oli havaittavissa tila, jossa normaalit tilanteet, kuten toisen kanssa vuorovaikutuksessa oleminen, ei tuonut henkilöille tyydytystä. Näillä henkilöillä oli usein vähemmän ystäviä palvelussa, he kommunikoivat toisten kanssa ja jakoivat kuvia harvemmin kuin muut. Jotkut tutkimukseen osallistujat piilottivat merkittäviä määriä tietoa profiileistaan ennen kuin he esittelivät niitä tutkijoille. Näissä henkilöissä oli havaittavissa merkkejä aistiharhasta, jossa henkilö kokee tapahtumien, joilla ei ole fyysistä syy-seuraussuhdetta, olevan jollain tavalla kausaalisia. Toimintojen salaaminen tutkijoilta saattoi olla myös merkki normaalia korkeammasta vainoharhaisuusasteesta. [33.]

Big datasta saattaisi olla merkittävää hyötyä myös pandemioiden tai laajojen epidemioiden ehkäisyssä. Jos terveydenhuollon tiedot voitaisiin yhdistää esimerkiksi tietullijärjestelmiin tai mobiililaitteiden paikkatietoihin, pystyttäisiin reaaliajassa seuraamaan sairauden leviämisenopeutta ja -suuntaa. Näiden tietojen avulla voitaisiin ennustaa seuraavia paikkakuntia, joissa epidemia on puhkeamassa ja toimittaa niihin valmiiksi esimerkiksi lääkintävarusteita tai väliaikaismajoituksia. Analyysien avulla leviämisen ehkäisyä voitaisiin tehostaa ja täsmentää esimerkiksi estämällä massapakoja sekä selvittää maantieteellisiä pisteitä, joista sairaudet ovat lähteneet liikkeelle.

7.5 Biotekniikka

Big data -teknologialla on merkittävä rooli genomitutkimuksen ja -analytiikan kehitymisessä. Ihmisen DNA koostuu kolmesta miljardista perusparista, mikä tekee tiedon käsittelystä haastavaa. DNA-tutkimus on tärkeää esimerkiksi syöpätutkimuksessa, mutta syövän ennakoimaton ja yksilöllinen kehittyminen sekä muuntautumiskyky vaikeuttavat hoitomenetelmien kehittämistä. Lisäksi jokainen syöpää sairastava henkilö tarvitsee yksilöityä hoitoa ja lääkitystä, sillä samat hoitomenetelmät eivät tehoa kaikkiin syöpätyyppeihin tai sovi kaikille henkilöille. Syövän analysointi edellyttää ihmisgenetiikan analysointia, mikä kuitenkin tarkoittaa myös tarvetta käsitellä valtavaa määrää tietoa. Tietoa on voitava käsitellä siten, että sen hyödyntäminen on tehokasta. Lisäksi tiedon tulee olla jäsennelty selkeästi, jotta tutkijat voivat käyttää sitä helpommin. DNA:n ja genetiikan bioteknisen tutkimuksen kustannukset ovat halventuneet merkittävästi viime vuosina, mikä on johtanut huomattavaan kasvuun tallennettavan tiedon määrässä. Suuri tiedon määrä laajentaa analyysimahdollisuuksia, mutta kääntöpuolena hallitse-

maton tietomäärän kasvu useista eri lähteistä saattaa koitua ongelmaksi ilman tehokkaita teknisiä menetelmiä. [34.]

Kaliforniassa 2011 perustettu biotekniikan tutkimusyritys Bina Technologies on kehittänyt menetelmän, jolla ihmisgenomi voidaan analysoida alle neljässä tunnissa. Yhdysvaltain hallituksen kansallisen ihmisgenomin tutkimuslaitoksen mukaan genomien sekvensoinnin kustannukset ovat pudonneet reilussa kymmenessä vuodessa 95,3 miljonnasta dollarista 6,6 tuhanteen dollariin. Kokonaisen genomien sekvensointi tuottaa puoli teratavua melko hajanaista ja jäsentelemätöntä tietoa, minkä käsittelyyn Bina Technologies on tuonut big data -analytiikkaa. Tämän teknologian myötä genomianalyysin hinta saadaan laskettua muutama sataan dollariin. Usein esimerkiksi vastasyntyneet, jotka sairastavat jotain geneettistä sairautta joutuvat viettämään sairaalassa pitkiä aikoja heidän tilansa diagnosoimiseksi. Nopea ja edullinen DNA-analyysi mahdollistaisi ongelman selvityksen siten, että tarvittavaa hoitoa voidaan antaa heti syntymän jälkeen. Nykyisten analyysimenetelmien kestosta johtuva kahden viikon odotus saattaa olla lapselle kohtalokasta. [35.]

7.6 Kiinteistöala

Kiinteistöalan asiantuntijan, diplomi-insinööri Ville-Petteri Riihisen mukaan kiinteistöliiketoiminnassa olisi merkittävää kysyntää tietojärjestelmälle, joka on kaikkien hyödynnettävissä ja joka tarjoaa informaatiota kiinteistöjen keskeisistä julkisista tiedoista. Tietopohjan rakentaminen nykyisillä teknologisilla menetelmillä on mahdollista, mutta mitään keskitettyä tietovarastoa tai valmiuksia markkinat kattavalle analysoinnille ei Suomessa ole vielä olemassa. Kaikki kiinteistöjä ja niillä sijaitsevia rakennuksia koskevat tiedot, kuten kaavoitusinformaatio, tiedot rakennuksista sekä omistus- ja vuokraustiedot ovat oleellisia kiinteistöjen myynti-, vuokraus- ja markkinointitoiminnoissa. Olisi myös kuluttajien edun mukaista, jos esimerkiksi kiinteistöjen huoltokustannukset, hintatiedot, vuokratasot, pinta-ala-, tilavuus- ja rakennuksien lupatiedot olisivat keskitetysti saatavilla. Kunnalliset rakennusvalvontavirastot ylläpitävät tietokantaa monista kiinteistöjä koskevista ominaisuuksista, mutta tarvetta on järjestelmälle, joka olisi kaikkien osapuolien käytössä. Tällaisella järjestelmällä olisi myös potentiaalia toimia kaupallisena tietolähteenä yrityksille, joiden liiketoiminnassa on liitántöjä kiinteistömarkkinoihin. Järjestelmään olisi hyvä lisätä esimerkiksi älykkään sähköverkon tai muun energiatehokkuutta

mittaavan järjestelmän tuottamaa tietoa, jotta analyysien tueksi saataisiin myös energiatehokkuudesta kertovaa tietoa. [36.]

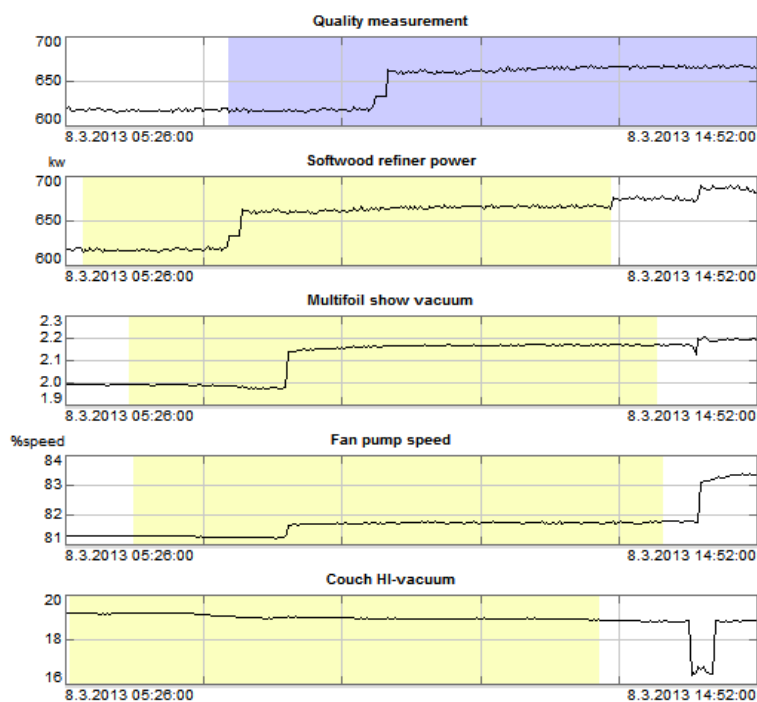
Verkossa toimiva kiinteistöjen välityspalvelu Trulia lanseerasi vuoden 2012 alussa big data -alustalla toimivan palvelun, joka auttaa kiinteistönvälittäjiä tunnistamaan potentiaaliset asiakkaat ja heidän asuntomieltymyksensä. Palvelu paljastaa henkilöt, joille on esimyönnetty lainaa ja jotka etsivät asuntoa. Lisäksi se tunnistaa näiden henkilöiden vaatimuksia uuden asunnon suhteen tehtyjen asuntohakujen perustella. Palvelun avulla saatetaan yhteen asiakas ja kiinteistönvälittäjä. Palvelun algoritmeissa on käytetty hyödyksi myös online-kyselyjä, hakukonekäyttäytymistä ja sähköpostiliikenteen kaavoja. [37.]

7.7 Teollisuus

Teollisuudessa on paljon aloja, joissa laajan tietopohjan analytiikkaa voidaan hyödyntää esimerkiksi toiminnan tehostamisessa ja vikatilanteiden selvityksessä. Yksi tiedon analysointiin perustuva järjestelmä on paperiteollisuuden käytössä oleva Savcor Wedge. Savcor Forest Group on suomalainen yritys, joka palvelee metsäteollisuutta tarjoamalla ratkaisuja laadun ja suorituskyvyn parantamiseen. Wedge on prosessianalyysituote, jolla monitoroidaan ja analysoidaan pääsääntöisesti paperiteollisuuden prosesseja. Järjestelmää voidaan hyödyntää minkä tahansa jatkuvaprosessisen toiminnan seurantaan myös muilla toimialoilla. Wedge-järjestelmän avulla käyttäjät pystyvät seuraamaan prosessien määrää ja laatua, sekä selvittämään niissä ilmeneviä ongelmia. Prosessin kulkua seurataan jatkuvasti. Kaikki prosessitapahtumat ja valvontainformaatio siirtyy keskitettyyn järjestelmään analysoitavaksi. Wedge-järjestelmän käyttöliittymän kautta käyttäjä pääsee porautumaan prosessin komponentteihin, joista ongelmien juurisytyt ovat havaittavissa. Järjestelmä tarjoaa reaaliaikaisesti tietoa prosessien eri vaiheista ja niiden kehityssuunnasta, jolloin tietoa voidaan käyttää päätöksenteon tukena. [38.]

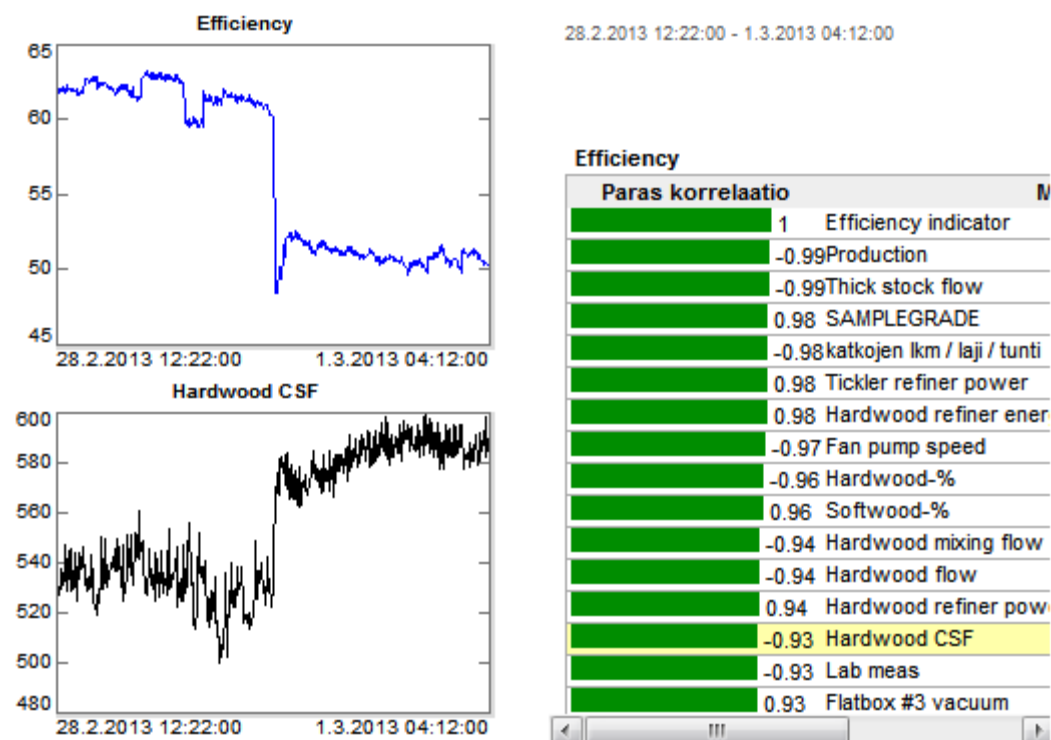
Wedge-järjestelmässä analysoidaan aikasarjoja, joissa havaintojen mittaus tapahtuu peräkkäisinä ajankohtina. Yhdestä mittauspisteestä voidaan vastaanottaa ainakin sata mittausta sekunnissa. Näitä sarjoja analysoimalla voidaan tutkia mittausmuutosten taustavaikuttajia. Analyysit eivät ole lähtökohtaisesti automaattisia, vaan käyttäjän on määriteltävä analyysin lähtökohdat. Analyysiä voidaan tuottaa reaaliaikaisesti virtaus-

prosessoimalla tai hyödyntämällä tallennettua informaatiota. Järjestelmällä on oma tallennusjärjestelmänsä, mutta lähtökohtaisesti käytetään kohdejärjestelmiin tallennettua tietoa. Jos kohdejärjestelmä ei tallenna mittaustuloksia, käytetään aina Wedge-järjestelmän aikasarjatietokantaa. Järjestelmään voidaan rakentaa jatkuvia mittareita, mutta pääroolissa on prosessin ongelmakohtien selvittäminen. Niitä voidaan havaita esimerkiksi mittaustuloksien äkkinäisistä muutoksista tai muutoin epätavallisista arvoista. Kuvassa 11 on nähtävissä esimerkki järjestelmän keräämästä tiedosta viidestä eri mittauspisteestä. [38.]



Kuva 11. Sensoreiden mittaustuloksia [38].

Tyypillinen ongelmaindikaattori graafisessa esityksessä on kuoppa tai pomppu, jolloin prosessi on todennäköisesti ollut keskeytynyt. Mikäli kyseessä on tiedostettu tapahtuma, se voidaan leikata analyysistä pois, jotta se ei vaikuta esimerkiksi kokonaistrendi-analyysiin. Porrasmaiset muutokset kertovat tavallisesti asetusarvojen muutoksista. Järjestelmä analysoi tapahtumien välisiä yhteyksiä ja ehdottaa todennäköisiä juurisyitä tutkitulle tapahtumakokonaisuudelle (kuva 12). Muutostilassa voidaan porautua syvemmälle aina yksittäiseen anturin lähettämään informaatioon saakka. [38.]



Kuva 12. Muutosanalyysiä antureiden tuottamasta tiedosta [38].

Savcor Wedge ei hyödynnä suoranaisesti big data -teknologiaa, sillä sen käyttökohteiden tuottaman tietomäärän analysointiin on riittänyt perinteiset teknologiaratkaisut. Analysointimenetelmät ovat kuitenkin vastaavia kuin big data -analytiikassa ja samankaltaisia algoritmeja voidaan hyödyntää laajempaankin analysointiin. Erot löytyvät prosessointitekniikasta ja analysointivolyymistä. Jos analyysijä haluttaisiin tuottaa samanaikaisesti merkittävästä määrästä tehtaita, tai jos analytiikkaan halutaan yhdistää prosessianalyysien lisäksi muutakin tietoa, big data -teknologia tullee ajankohtaiseksi. Tällaisia muita tietolähteitä voisivat olla esimerkiksi kulunvalvonta, energiankulutukseen liittyvät mittarit, liiketoiminnan kannattavuusanalyysit tai meteorologisten tapahtumien vaikutukset raaka-aineiden tuotantoon.

8 Mahdollisuudet ja haasteet

Big data -teknologia tuo paljon uusia mahdollisuuksia toiminnan tehostamiseen ja palveluiden kehittämiseen. Uudet innovaatiot ja internetin käytön lisääntyminen ovat saaneet ihmiset jättämään itsestään yhä enemmän digitaalisia jälkiä. Tietoyhteiskunnassa yksityisyyden rajat kuitenkin hälvenevät ja ihmisten haavoittuvuus kasvaa. Yksilöiden

valvonta tehostuu, mutta riskinä on kohtuuden tai tunkeilevuuden rajan ylittyminen. Esimerkiksi internetissä tapahtuvat toiminnot kertovat paljon henkilöstä.

Etenkin sosiaalisessa mediassa tapahtuva toiminta saattaa paljastaa henkilöstä merkittävästi tietoa, kuten selviää yhdysvaltalaisen PNAS-tiedelehden tutkimusjulkaisusta. Tutkimuksen mukaan Facebookin tykkää-napin painalluksista voidaan päätellä esimerkiksi henkilön uskonnollinen kanta tai seksuaalinen suuntautuminen. Tutkimus osoitti myös, että napin painallus kertoo 95 prosentin tarkkuudella henkilön ihonvärin, 85 prosentin tarkkuudella poliittisen suuntautumisen ja 65 prosentin tarkkuudella sen, käyttääkö henkilö huumeita. [39.]

Ihmiset ovat tottuneet elämään teknologisesti hienostuneessa ympäristössä, jossa pienet arjen ongelmat ovat helposti tekniikalla ratkottavissa. Kotitalouksissa älykkäät kodinkoneet yleistyvät jatkuvasti vähentäen käyttäjän roolia niiden käytössä ja esimerkiksi modernia teknologiaa hyödyntävät elektroniset kauppalistat muistuttavat, jos jotain jäi ostamatta. Sen lisäksi, että nämä tekniset välineet helpottavat käyttäjän arkea, niiden avulla voidaan myös kerätä tietoa uusien toimintamallien, tuotteiden ja palveluiden kehittämiseksi. Keräämällä tietoa nykytuotteiden käytöstä ja käytön ominaispiirteistä, voidaan tuotteiden seuraava sukupolvi valmistaa vastaamaan entistä paremmin käyttäjien tarpeita.

Populaation segmentointi on avainasemassa, kun halutaan kohdentaa markkinoita, räätälöidä tuotteita tai palveluita. Liian pitkälle viety analysointi ja sen hyödyntäminen tuovat kuitenkin eettisiä haasteita mm. kohdennetun markkinoinnin ratkaisuihin. Esimerkiksi tulevaisuudessa äidiksi analysoitu nainen on voinut saada keskenmenon. Minkään yrityksen imagolle ei tekisi hyvää markkinoida lastenhoitovälineitä tällaiselle henkilölle, tai esimerkiksi tarjota vanhustenhoidollisia palveluita juuri menehtyneen henkilön lähiomaisille [18, s. 90].

Yksityisyyden rajojen hälveneminen on yksi suurimpia big datan haasteita. Tosin sukupolvi sukupolvelta asenteet ja käsitykset yksityisyyden rajoista ovat muuttumassa. Suuret ikäluokat ovat usein hyvinkin tarkkoja varjelemaan omaa yksityisyyttään, kun taas nuoret henkilöt ovat tottuneet tuomaan lähes koko elämänsä esille sosiaalisen median palveluissa, kuten Facebookissa, Twitterissä tai LinkedInissä. Nuorempi sukupolvi ei enää pidä vastaavalla tavalla yksityisyyttä loukkaavina, ei-hyväksyttävänä tai epäilyttävänä samoja asioita kuin heidän vanhempansa tai isovanhempansa. Internetissä jul-

kaistu kuva henkilöstä juhlimassa ei välttämättä enää ole osoitus henkilön kykenemättömydestä kantamaan vastuuta työtehtävistään, eikä julkisesti ilmaistuja, valtavirrasta poikkeavia mielipiteitä pidetä maanpetturuutena vaan enemmänkin rehellisyytenä ja rohkeutena olla oma itsensä. Entisaikojen tabut, kuten mielen sairaudet, seksuaaliseen vähemmistöön kuuluminen tai muunlainen poikkeavuus valtaväestöstä, nostetaan usein esille jo senkin takia, että näille ilmiöille halutaan antaa kasvot. Harva nuori on nykyään edes tietoinen esimerkiksi siitä, ettei naisten ollut vielä muutama vuosikymmen sitten suotavaa mennä yksinään ravintolaan.

Yksityisyyden menettämisen uhkia on useita eri tyyppejä aina ärsyttävästä mainonnasta syrjintään saakka. Kuinka tasapainotella sen välillä, että big data -konseptista saadaan kaikki hyöty irti, mutta samalla vältetään sitä kaikkea mahdollista harmia mitä tiedon keruusta, jakamisesta ja analysoinnista saattaa aiheutua? Läpinäkyvyys ja tiedon tuominen kaikkien ulottuville parhaimmillaan vähentävät huomattavasti tiedon haku- ja prosessointiaikaa, jolloin toimenpiteiden vasteaika lyhenee ja tuotettavan palvelun laatu paranee. Äärimmäisen sensitiivistä tai muuta vastaavaa informaatiota on kuitenkin suojeltava, jotta kansallista turvallisuutta tai ihmishenkiä ei saatettaisi vaaraan. Henkilökohtaiset tiedot, kuten terveys- ja varallisuustiedot ovat usein niitä, joiden analysoinnista saatava hyöty on yksilölle merkittävintä. Vastaavasti niiden käyttö herättää myös paljon kysymyksiä, kuten kuka tiedon omistaa, kenellä on oikeudet sillä saavutettavaan hyötyyn, mitä oikeuksia on tällaisen tiedon käyttämiseen ja kuka on vastuussa, jos virheellinen tieto aiheuttaa negatiivisia seuraamuksia. [1, s. 11–12.]

Tiedon käyttö ei ole ainoa haasteellinen sektori, johon tulee kiinnittää huomiota. Tiedon keruu-, säilytys- ja siirtosäännöksiä tarvitaan ainakin tilapäisesti, sillä teknologia mahdollistaa hyvinkin arkaluonteisten tietojen tallennuksen. Näitä ovat esimerkiksi luottokorttitiedot, paikkatiedot, terveydenhuollolliset tiedot ja henkilökohtaiset viestintätapahtumat mukaan lukien niiden sisältö. Näiden tietojen perusteella voidaan tehdä hyvin tarkkoja olettamuksia yksilön mielipiteistä, sosiaalisesta toiminnasta sekä taloudellisesta, henkisestä ja fyysisestä tilasta. Tällaisilla analyyseillä voidaan kuitenkin tehdä johtopäätöksiä myös tulevista päätöksistä, esimerkiksi harkinnassa olevista hankinnoista aina rikollisiin aikeisiin saakka. Viranomaisilla yksi todennäköisimmistä käyttötarkoituksista onkin taistelu rikollisuutta vastaan ehkäisemällä niitä jo ennen tapahtumahetkeä.

Kaupallinen fokus on taas siinä, että erilaisia kuluttajia kohdellaan eri tavalla. Personoidut mainokset ja tarjoukset, asiakaskohtaiset hinnat tai erilaiset päätökset luoton,

vakuutuksien tai työn suhteen voisivat perustua henkilöiden syvempään analysointiin. Ajatus on houkutteleva, mutta ylilyönnit ovat mahdollisia. Haluammeko todella elää maailmassa, jossa kaikki toimii niiden indikaattorien mukaan, jotka on analysoitu fyysisestä terveydestä, henkisestä tasapainosta, seksuaalisista mieltymyksistä tai avioinnesta? Jos eläisimmekin sellaisessa maailmassa, harkitsisimmeko jokaista liikettä niin internetkäyttäytymisessä kuin hankinnoissa peläten leimatuksi tulemista niiden perusteella? [18, s. 84–86.]

Näitä pelkoja ja haasteita voitaisiin kuitenkin kääntää vahvuudeksi esimerkiksi nostamalla mielenterveysongelmien tunnistusastetta. Sosiaalisen median tapahtumiin voitaisiin kytkeä automaattisia analyysijärjestelmiä ja yhdistää niihin terveydenhuollon keräämää tietoa ja esimerkiksi työpaikan kulunvalvonnan tietoja. Nämä yhdessä internetin hakuhistorian tai paikkatietojen kanssa saattaisivat paljastaa riskitapauksia ennen kuin henkilö ehtii vahingoittamaan itseään tai muita. Teknologia ja tiede tulevat varmasti mahdollistamaan tämänkaltaisen järjestelmän, mutta yksityisyys ja yksilön oikeudet tulevat olemaan suurimmat esteet sen käyttöönotolle. Monet mielenterveyshäiriöt, syrjäytyminen ja päihderiippuvuudet jäävät kuitenkin nyky menetelmillä tunnistamatta. Ratkaisun voisi tarjota eräänlainen avoin luokittelujärjestelmä. Yksilöllä olisi aina mahdollisuus saada tieto siitä, mitä analyysijärjestelmästä hänestä on tehty tai mitä tietoja hänestä on tallennettu. Diagnoosialgoritmien perusteella henkilöt luokiteltaisiin anonyymisti riskiluokkiin ja henkilölle itselleen tarjottaisiin automaattisesti tietoa ja suosituksia apukeinoista tai mahdollisista hoitomenetelmistä. Jos kuitenkin luokitus nousisi tietyn hälytysrajan yli, tieto siitä menisi viranomaistaholle jatkotoimien harkintaa varten ja kohde identifioitaisiin. Esimerkiksi jos analyysien perusteella voitaisiin arvella henkilön mahdollisesti sairastavan kaksisuuntaista mielialahäiriötä, hänelle voitaisiin automaattisesti toimittaa tietoa sairaudesta. Henkilö voisi tällöin itse harkita, kokeeko hän sairauden elämänlaatu huonontavaksi ja haluaako hän jatkohoitoa. Jos taas analyysien perusteella voitaisiin arvella henkilön sairastavan pahanlaatuista paranoidia skitsofreniaa ja hakevan internetistä tietoa massatuhoaseista, ihannoivan koulusurmaajia ja verkostoituvan muiden korkean riskin henkilöiden kanssa, tilannetta voitaisiin tutkia, seurata ja tarvittaessa siihen voitaisiin puuttua pakkokeinoin.

Big datalla on mahdollisuus tarjota syöpähoitotutkimuksille tutkimuksia tehostavia apumenetelmiä, sillä syövän geneettisen perustan selvittäminen edellyttää yli tuhannen genomin analysointia. Tämä ei ole aiemmilla teknologioilla ollut mahdollista, mutta jo nyt on havaittavissa esimerkkejä siitä, kuinka kansalliset säädökset kuten Yhdysval-

loissa terveydenhuoltoa koskeva yksityisyysmääräys HIPAA rajoittavat jatkuvasti tutkimuksien etenemistä. Yksityisyysmääritykset rajoittavat jatkuvasti lääketieteellisen tiedon jakoa ja käyttöä, jonka vuoksi tappavien sairauksien tutkimukset viivästyvät tai niitä jopa perutaan tutkijoiden ollessa aseettomia vaatimuksia vastaan. Esimerkiksi sydän-sairauksien Minnesota Heart Study -tutkimusohjelmaan kuuluvat lääkärit kävivät aiemmin läpi kaikki kaupungin sydänperäiset ensiapupoliklinikan käynnit viiden vuoden välein löytääkseen ja raportoidakseen hoitomenetelmiä sekä niiden vaikutuksia. Tämä yksi maailman tärkeimmistä sydäntutkimuksista jouduttiin keskeyttämään, sillä HIPAA edellytti potilassuostumusta tietojen käsittelyyn. Tutkimusohjelmalla ei ollut resursseja jäljittää potilaita jälkikäteen eikä suostumuksen saanti sydänkohtauksen hetkellä ollut aina mahdollista. [18, s. 84–86; 40.]

Haasteita lääketieteen sektorilla aiheuttaa myös globaalien sopimuksien puuttuminen informaation ja bioteknisten näytteiden jakamisesta. Lisäksi suurin osa potilastiedoista on edelleen tallennettu paperiformaattiin, josta sitä ei saa helposti siirrettyä tarvittavaan digitaaliseen muotoon. Tätä on jo lähdetty muuttamaan kannustamalla hoitohenkilökuntaa käyttämään elektronisia tallennusmenetelmiä, vaikka tietoturvallisuus ja intimitetisuoja ovat tuoneet digitalisointiin lisäkustannuksia. Ongelmana on myös, että sellaisten henkilöiden vähyys, jotka pystyvät analysoimaan lääketieteellistä tietoa ja joilla on riittävät tietotekniset taidot. [18, s. 66–68.]

Riittävän laajan ammattitaidon puute on suuri ongelma myös muilla big datan sovellusalueilla. Päätöksentekoa voitaisiin tukea tai automatisoida big data-algoritmeilla, joiden tuottamat analyysit ovat riittävän hienostuneita minimoimaan riskejä ja nostamaan piiloarvot esille. Toistaiseksi sellaisia henkilöitä, joilla on kyky tuottaa tällaisia algoritmeja ja joilla on tutkittavasta aihealueesta vankka asiantuntemus, ei suoranaisesti kouluteta missään. Jotta analyysiteknologian käyttöaste saadaan korkeaksi, on tehtävä muutoksia myös koulutusrakenteeseen. Koulutusrakenteen muutokset taas edellyttävät rakenteesta vastaavien tahojen painostusta ja tarpeen esille tuontia. Toistaiseksi organisaatioiden johdolla ei ole vielä riittävää ymmärrystä big datasta, jotta tällaisia vaatimuksia osattaisiin esittää.

Esimerkkinä kahvilaketju Starbucks on käyttänyt asiakaskortteja asiakastiedon keruuseen. Yrityksellä on asiakastietoa runsaasti, mutta sen hyödyntäminen on ollut pienimuotoista. Yrityksen johtoa onkin yhtiön sisällä syytetty välinpitämättömyydestä tiedon hyödyntämistä kohtaan. Starbucksin liiketoiminta-analytiikasta vastaavan johtajan, Joe

Cugnan mukaan kuuden miljoonan korttikäyttäjän kaikki tiedot ovat tallessa ja he tietävät, kuinka kukin käyttäjä toisistaan eroaa, mutta suoranaisesti liiketoimintaa hyödyntävää käyttöä yhtiö ei tiedoille ole onnistunut löytämään [41].

Koska organisaatioilla ei ole vielä riittävää osaamista tietomassojen tulkitsemiseksi, työnkulkua tai prosesseja ei suunnitella siten, että big datan hyötykäyttö olisi optimoitua. Organisaatiot tarvitsevat jatkuvasti uusia tietojärjestelmiä ja niiden integroiminen olemassa oleviin järjestelmiin on oleellista. Vastaavasti tulisi ottaa huomioon myös näiden tietojärjestelmien hyödynnettävyys tiedon analysoinnissa. Tämä aiheuttaa tarpeita hankinnoille ja projekteille, joissa tietoa yhdenmukaistetaan. Näiden investointien on oltava selitettävissä johdolle uudesta teknologiakäsitteestä huolimatta, sillä big data -analysoinnissa tietoa pitää kerätä useista eri lähteistä. Koska tietoa joudutaan mahdollisesti ostamaan myös organisaation ulkopuolelta, tiedon yhtenäistämismahdollisuus on oleellista. Etenkin kansainvälisessä tiedonvaihdoissa on huomioitava erilaiset kansainväliset standardit, sillä ne saattavat aiheuttaa tilanteen, jossa tietoja ei voida yhdistää ja analysoida sellaisenaan. Esimerkiksi Yhdysvalloissa on tapana ilmaista lukua tuhat muodossa 1,000, mikä taas on Suomessa tulkittavissa luvuksi yksi kolmen desimaalin tarkkuudella. Tietomarkkinoilla tarjolla olevien tietojen lähteet eivät myöskään ole välttämättä punninneet tiedon jakamista laajalti. Lisäksi tällaiset tiedot saattavat olla merkittävässä roolissa tuomaan kilpailuetua markkinoilla, jolloin sitä ei edes haluta jakaa kilpailijoiden kanssa. Tällöin on harkittava riittävää kompensatiota tiedon käytöstä

Tiedon hallinnassa on mietittävä sen sijoitusta ja sijoituksen merkitystä sovellettavan lainsäädännön kannalta. Esimerkiksi pilviteknologiassa tiedot saattavat fyysisesti sijaita missä tahansa, kuten myös Hadoop-klusterille rakennetuissa ratkaisuissa. Kaikki valtiot eivät ole vastuuntuntoisia ja välittäviä, jolloin tiedon joutuminen heidän käsiinsä saattaa aiheuttaa kansainvälisiä riskejä. Tällaiset valtiot saattavat itsekin hyödyntää tiedonkeruuteknologiaa saadakseen lisätyökaluja mm. kansansortoon. New Yorkin yliopiston opiskelija Drew Conway teki Wikileaks-materiaalista analyysiä Afganistanin konflikteista ja niiden kehityssuunnista. Hän jakoi materiaalin alueittain ja kategorisoi sen tyyppin vihamieliseen, neutraaliin ja ystävälliseen tunnistaakseen tapahtumien kaavoja. Näiden tietojen perusteella voitiin vahvistaa joukkojen aktiivisuuden teorioita ja löytää sesonki-
piikkejä konfliktitilanteista Taliban-joukkojen kanssa sekä paikantaa konfliktit tietyille alueille [42]. Tämänkaltaiset analyysit saattaisivat olla väärissä käsissä kohtalokkaita,

mikäli ne laajennettaisiin koskemaan myös muita sotilaallisia tapahtumia kuin pelkäämään konfliktitilanteita.

9 Yhteenveto

Tämän insinööriyön tarkoituksena oli tutustua uuteen tiedonhallinnan ja analysoinnin konseptiin, big dataan ja tutkia sen mukanaan tuomia hyötyjä, mahdollisuuksia ja haasteita. Big data on vasta nouseva teknologiakokonaisuus, eikä sen kaikesta potentiaalista ole vielä laajoja käytännön esimerkkejä, vaikka etenkin suuremmat teknologiayritykset pyrkivätkin markkinoille omilla big data -sovelluksillaan. Saatavilla oleva materiaali on pääasiassa Yhdysvalloissa tuotettua, tutkimuslaitosten julkaisuja tai mediassa julkaistuja artikkeleita. Ongelmalliseksi laajamittaisten julkaisujen tuottamisen tekee myös se, että analytiikassa on yleensä kyse kertaluonteisesta tai yksilöllisestä tarpeesta. Kehitetyt algoritmit ja analyysimenetelmät halutaan pitää kilpailijoilta salassa, koska liiketoimintaan liittyvän yksityiskohtaisen tiedon tai sitä koskevien analyysien julkaisu saattaisi olla liikkeenharjoittajalle todella haitallista.

Työn tuloksena syntyi kokonaisuus, joka kiteyttää big datan ideologian ja komponentit sekä tuo esille sen hyödyntämistapoja ja potentiaalia. Jo pelkäämään big data -alustan toteutuksen teknisiä vaihtoehtoja on markkinoilla runsaasti, mutta hyödyntämismahdollisuuksia on olemassa sitäkin enemmän. Tämä työ tarjoaa lisätietoa henkilölle, jotka ovat kiinnostuneita aiheesta mutta joilla ei ole siitä vielä riittävää pohjatietoa. Big data on nouseva teknologia, josta ei ole vielä olemassa riittävästi yleisen tason materiaalia. Tarjolla oleva tieto koostuu pääsääntöisesti erillisistä, jonkin osa-alueen dokumenteista ja artikkeleista. Tässä selvityksessä on haluttu tuottaa suomenkielinen opas big datasta ja koostaa hajanaisista tietolähteistä kokonaisuus, joka on kaikkien ymmärrettävissä.

Työstäessäni opinnäytetyötä huomasin kiinnostäväni entistä enemmän huomiota esimerkiksi internet-sivujen kehitysissä esiintyviin mainoksiin ja hakukoneen sivuehdotuksiin. Todellisuus siitä, kuinka paljon eri tahot keräävät dataa, analysoivat minua ja profiloivat verkkokäytöstäni yksilöityä markkinointia varten, yllätti minut todenteolla. Alkaessani kirjoittaa työtä aiheesta, jonka koin samalla sekä mielenkiintoiseksi että hyödylliseksi, en vielä ymmärtänyt, kuinka laaja kokonaisuus oli kyseessä. Uskon, että big data -konseptissa on tulevaisuus, kunhan sen ydinajatus ymmärretään ja kokonaisuutta opitaan soveltamaan.

Teknologian kehitys on aina kohdannut merkittävää muutosvastarintaa. On kyse sitten ollut uuden energiamuodon kehittämisestä tai sosiaalisen median sovelluksien toimintaperiaatteiden muutoksista. Yksityisyyden sorkkiminen nykYTEknologian mahdollistamalla intensiteetillä tulee tuottamaan päänvaivaa niin valtioille ja niiden eri sektoreille, yrityksille kuin yksiköllekin. Sukupolvelta sukupolvelle yksilöiden asenteet kuitenkin muuttuvat yhä avoimemmiksi, mikä mahdollistaa entistä laajemman tallennetun tiedon hyötykäytön. Muutos näyttää väistämättömältä. Ehkä meidän tulisikin ottaa rooli teknologian kehittämisessä ja tukea teknologian kasvua sen sijaan, että pitäisimme kiinni vanhasta tai pyrkisimme kasvamaan teknologiakehityksen mukana. Kannatan ajatusta, jossa resurssit ja asiantuntemus kanavoidaan vastustamisen sijaan siihen, että teknologiasta saadaan kaikki hyöty irti mahdollisimman miellyttävällä tavalla. Kun kehitämme itse, toiset eivät tee sitä puolestamme.

Lähteet

- 1 McKinsey Global Institute. 2011. Big data: The next frontier for innovation, competition and productivity.
- 2 Security Information and Event Management. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Security_information_and_event_management. Luettu 29.1.2013.
- 3 Business Intelligence. 2013. Verkkodokumentti. Wikipedia. http://fi.wikipedia.org/wiki/Business_intelligence. Luettu 15.3.2013.
- 4 The 2011 Digital Universe Study: Extracting Value from Chaos. Verkkodokumentti. 2011. <http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm>. Luettu 11.3.2013.
- 5 Big Data. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Big_data. Luettu 14.2.2013
- 6 Pervilä, Markku. 2012. Big data luo miljoonia it-työpaikkoja. Verkkodokumentti. http://www.tietoviikko.fi/cio/big+data+luo+miljoonia+ittyopaikkoja/a850520?s=bu_tekta. 25.10.2012. Luettu 14.2.2013.
- 7 DeLua, Julianna. 2011. Big Data Unleashed Part 3: Five considerations for your information management agenda with big data. Verkkodokumentti. <http://blogs.informatica.com/perspectives/2011/07/05/big-data-unleashed-part-3-five-considerations-for-your-information-management-agenda-with-big-data>. 5.7.2011. Luettu 14.2.2013.
- 8 Fujitsu Finland. 2013. The Cloud Office's Newsletter. Yrityksen sisäinen dokumentti. Luettu 15.3.2013.
- 9 Apache Hadoop. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Apache_Hadoop. Luettu 29.1.2013.
- 10 MapReduce. 2013. Verkkodokumentti. Wikipedia. <http://en.wikipedia.org/wiki/MapReduce>. Luettu 29.1.2013.
- 11 In-Memory Processing. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/In-Memory_Processing. Luettu 26.3.2013
- 12 MongoDB. 2013. Verkkodokumentti. Wikipedia. <http://en.wikipedia.org/wiki/MongoDB>. Luettu 27.3.2013.

- 13 SIMD. 2013. Verkkodokumentti. Wikipedia. <http://fi.wikipedia.org/wiki/SIMD>. Luettu 27.3.2013.
- 14 Marz, Nathan. 2012. Storm tutorial. Verkkodokumentti. <https://github.com/nathanmarz/storm/wiki/Tutorial>. Luettu 27.3.2013.
- 15 Woods, Dan. 2012. Ten properties of the perfect big data storage architecture. Verkkodokumentti. <http://www.forbes.com/sites/danwoods/2012/07/23/ten-properties-of-the-perfect-big-data-storage-architecture>. 23.7.2012. Luettu 1.3.2013.
- 16 Pilvilaskenta. 2013. Verkkodokumentti. Wikipedia. <http://fi.wikipedia.org/wiki/Pilvilaskenta>. Luettu 18.3.2013.
- 17 Big data pilvipalvelut. 2013. Verkkodokumentti. <http://hadoopbigdata.wordpress.com/pilvipalvelut>. Luettu 18.3.2013.
- 18 GigaOM Pro. 2012. A near-term outlook for big data.
- 19 Mlovett. 2011. Machine-to-Machine Technology = Efficient Economy. Verkkodokumentti. <http://blog.trentonsystems.com/machine-to-machine-technology-efficient-economy>. 11.11.2011. Luettu 17.3.2013.
- 20 Hochanadel, Kurt. 2010. Providing Data in Real Time: Machine to Machine Systems Smooth Transportation. Verkkodokumentti. <http://rtcmagazine.com/articles/view/101554#>. Luettu 17.3.2013.
- 21 Horizon Energy Research. 2013. Probing responses to future smart grid technologies. Verkkodokumentti. <http://horizonenergy.blogspot.fi>. Luettu 17.3.2013.
- 22 Smart grid. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Smart_grid. Luettu 14.2.2013.
- 23 Meter Data Management. 2012. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Meter_data_management. Luettu 14.2.2013.
- 24 Operational Intelligence. 2013. Verkkodokumentti. Wikipedia. http://en.wikipedia.org/wiki/Operational_intelligence. Luettu 18.3.2013.
- 25 Duhigg, Charles. 2012. How companies learn your secrets. Verkkodokumentti. <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&r=2&hp&>. 16.2.2012. Luettu 29.1.2013.
- 26 Hill, Kashmir. 2012. How Target figured out that a teen girl was pregnant before her father did. Verkkodokumentti.

- <http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did>. 16.2.2012. Luettu 29.1.2013.
- 27 Argillander, Timo. 2012. Miksi iso data on iso juttu?. Verkkodokumentti. <http://www.digitalmedia.fi/miksi-iso-data-on-iso-juttu>. 26.10.2012. Luettu 13.2.2013.
 - 28 Carr, David. 2013. Giving viewers what they want. Verkkodokumentti. http://www.nytimes.com/2013/02/25/business/media/for-house-of-cards-using-big-data-to-guarantee-its-popularity.html?pagewanted=all&_r=1&. 24.2.2013. Luettu 2.3.2013.
 - 29 Fehrenbacher, Katie. 2011. The Internet of Things and energy. Verkkodokumentti. <http://gigaom.com/2011/10/10/the-internet-of-things-energy>. 10.10.2011. Luettu 13.2.2013.
 - 30 Älykäs verkko eli Smart Grid. 2013. Verkkodokumentti. Energiateollisuus. <http://energia.fi/sahkomarkkinat/sahkoverkko/alykas-verkko>. Luettu 1.4.2013.
 - 31 Etäluettava sähkömittari. 2013 Verkkodokumentti. Wikipedia. http://fi.wikipedia.org/wiki/Et%C3%A4luettava_s%C3%A4hk%C3%B6mittari. Luettu 1.4.2013.
 - 32 Big data use case: Better Health Care, Lower Cost. 2013. Verkkodokumentti. Greenplum. <http://www.greenplum.com/industry-buzz/big-data-use-cases/better-health-care-lower-cost>. Luettu 15.3.2013.
 - 33 Wall, Timothy. 2013. Facebook Activity Reveals Clues to Mental Illness says MU Researcher. Verkkodokumentti. <http://munews.missouri.edu/news-releases/2013/0124-facebook-activity-reveals-clues-to-mental-illness-says-mu-researcher>. 24.1.2013. Luettu 15.3.2013.
 - 34 Finkelstein, Masha. 2012. Big Data in Genomics and Cancer Treatment. Verkkodokumentti. <http://hortonworks.com/blog/big-data-in-genomics-and-cancer-treatment>. 22.6.2012. Luettu 15.3.2013.
 - 35 Versel, Neil. 2013. Big data startup eyes genome analysis in 4 hours. Verkkodokumentti. <http://www.informationweek.com/healthcare/leadership/big-data-startup-eyes-genome-analysis-in/240148802>. 19.2.2013. Luettu 15.3.2013.
 - 36 Riihinen, Ville-Petteri. 2013. Diplomi-insinööri. Newsec Advice Oy associate. Haastattelu 22.3.2013.
 - 37 Barr, Alistair. 2012. Real estate site Trulia taps “big data” for new service. Verkkodokumentti. <http://www.reuters.com/article/2012/01/12/us-trulia-bigdata-idUSTRE80B0RD20120112>. 12.1.2012. Luettu 1.4.2013.

- 38 Kahala, Jarmo. 2013. Vice President, Process Analysis & Biorefineries, Savcor Forest Group. Haastattelu 28.3.2013.
- 39 Hallamaa, Teemu. 2013.
http://yle.fi/uutiset/facebookissa_tykkaaminen_kertoo_paljon_kayttajasta/653340
3. Verkkodokumentti. 12.3.2013. Luettu 13.3.2013.
- 40 Gomes, Lee. 2009. The Hidden Cost of Privacy. Verkkodokumentti.
http://www.forbes.com/forbes/2009/0608/034-privacy-research-hidden-cost-of-privacy_2.html. 20.5.2009. Luettu 13.3.2013.
- 41 Storås, Niclas. 2013. Starbucksilla on big data -ongelma: sitä on liikaa eikä johto ymmärrä. Verkkodokumentti.
http://www.tietoviikko.fi/kaikki_uutiset/starbucksilla+on+big+data+ongelma+sita+on+liikaa+eika+johto+ymmarra/a889712. 26.3.2013. Luettu 27.3.2013.
- 42 Smith, David. 2011. 5 real-world uses of big data. Verkkodokumentti.
<http://gigaom.com/2011/07/17/5-real-world-uses-of-big-data>. 17.7.2011. Luettu 1.4.2013.

Liite 1. Esimerkki Windows-järjestelmän kirjautumislokista epäonnistuneen kirjautumisyrityksen tapauksessa

2/19/13

8:36:55.000 AM

02/19/2013 08:36:55 AM

LogName=Security

SourceName=Microsoft Windows security auditing.

EventCode=4625

EventType=0

Type=Information

ComputerName=xxxx

TaskCategory=Logon

OpCode=Info

RecordNumber=44240

Keywords=Audit Failure

Message=An account failed to log on.

Account For Which Logon Failed:

Security ID:	NULL SID
Account Name:	<i>username</i>
Account Domain:	<i>domainname</i>

Failure Information:

Failure Reason:	Unknown user name or bad password.
-----------------	------------------------------------

Status:	0xc000006d
---------	------------

Sub Status:	0xc0000064
-------------	------------

Process Information:

Caller Process ID:	0x0
--------------------	-----

Caller Process Name:	-
----------------------	---

Network Information:

Workstation Name:	<i>computername</i>
-------------------	---------------------

Source Network Address:	-
-------------------------	---

Source Port:	-
--------------	---

Detailed Authentication Information:

Logon Process:	NtLmSsp
----------------	---------

Authentication Package:	NTLM
-------------------------	------

Transited Services:	-
---------------------	---

Package Name (NTLM only):	-
---------------------------	---

Key Length:	0
-------------	---