



BIG DATA JA YRITYKSEN MARKKINOINTI

Pekka Perolainen

Opinnäytetyö
Huhtikuu 2014
Tietojenkäsittely
Aikuisopiskelija

TAMPEREEN AMMATTIKORKEAKOULU
Tampere University of Applied Sciences

TIIVISTELMÄ

Tampereen ammattikorkeakoulu
Tietojenkäsittely
Aikuisopinnot

PEROLAINEN, PEKKA
Big data ja yrityksen markkinointi

Opinnäytetyö 40 sivua
Huhtikuu 2014

Opinnäytetyössä oli tavoitteena tutkia big datan hyödyntämistä yrityksen myyntityössä ja markkinoinnissa. Yrityksillä on mahdollisuuksia käyttää omista tai ulkoisista lähteistä kerättyä tietoa toimintansa tehostamiseen. Yrityksen omat tiedot ovat lähinnä transaktiotietoja, asiakaskorttitietoa, logistiikkadataa tai anturidataa. Kameratallenteet ovat myös osa yritysten keräämää dataa, lainsäädännössä tämä data lasketaan henkilökisteritiedoksi. Yritysten on mahdollista kerätä, käsitellä ja yhdistellä keräämäänsä tietoa, kun se täyttää lainsäädännössä tiedon käsittelylle määritellyt asiat. Kerättyjen tietojen käytölle on lisäksi eettisiä ja lainsäädännöllisiä rajoitteita.

Big datan hyödyntäminen markkinoinnissa perustuu tehokkaisiin analysointimenetelmiin. Big datan analysointia voidaan tehdä esimerkiksi Hadoop-pohjaisella tietojen rinnakkaiseen käsittelyyn perustuvalla järjestelmällä. Rinnakkaiseen käsittelymalliin perustuvasta Hadoop-pohjaisesta tietojenkäsittelystä on opinnäytetyössä suppea toimintakuvaus. Big datan hallintaan ja analysointiin Hadoop-järjestelmällä on kehitetty omia ohjelmistoja. Näistä ohjelmistoista on tähän opinnäytetyöhön kerätty lyhyt katsaus.

Euroopan unionin henkilötietodirektiivi ja henkilötietolaki ohjaavat tietojen tallennusta ja käsittelyä Suomessa. Nämä ohjaavat henkilökisterien muodostamista ja tietojen hallintaa. Henkilötietojen käsittelylle tai tietojen siirtämiselle Euroopan talousalueen ulkopuolelle on olemassa tarkat säädökset. Sosiaalisen median palveluja käyttäessään yksityinen henkilö tuottaa runsaasti henkilökisteriksi luettavaa tietoa. Kun asiakas liittyy yrityksen asiakasrekisteriin, kaikki hänen antamansa tiedot ovat henkilökisteritietoja. Omien tietojen julkituomisessa on palvelujen käyttäjän tunnettava oma vastuunsa.

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree Programme in Business Information Systems

PEROLAINEN, PEKKA
Big Data and Retail Marketing

Bachelor's thesis 40 pages
April 2014

The goal of the thesis was to study the use of big data from the perspective of marketing and sales in Finland. Commercial enterprises have the opportunity to use their own information or information collected from external sources to improve operational efficiency. Enterprises are allowed to collect, process and combine data, when the data processing meets the Finnish legal regulations.

Big data is based on the use of efficient analysis methods. The analysis can be done, for example, using a Hadoop-based parallel information processing system. Around the open source Hadoop, many companies have developed their own software for big data management and analysis.

The European Union data protection directive and Finnish laws based on the directive regulate the use and processing of personal data in Finland. Legislation directs and supervises the formation and management of personal data in Finland and within the European Union. Personal data may only be transferred to third countries if that country provides an adequate level of protection. A single person produces personal data file information by using social media services. Also companies collect personal data for their customer registers. The responsibility for controlling how much personal information is given is left to the person who provides that information about him/herself.

Keywords: big data, marketing, data privacy

SISÄLLYS

1	JOHDANTO.....	7
2	BIG DATASTA HYÖTYÄ YRITYKSELLE	8
2.1	Big data	8
2.2	Yrityksen oma liiketoimintatieto	10
2.3	Kuluttajan tuottama tieto yritykselle.....	11
2.4	Datan jatkojalostus uusilla analysoinneilla.....	11
3	MENETELMÄT BIG DATAN HYÖDYNTÄMISEEN	14
3.1	Mistä tietoa yritykselle.....	14
3.2	Yksityinen eli suljettu tieto yrityksessä	14
3.3	Julkinen avoin tieto eli open data yrityksen saatavilla	15
3.4	Data-analyysi	16
3.5	Tiedosta liiketoimintaa	17
4	BIG DATAN ANALYSOINTIMENETELMIÄ.....	19
4.1	Tietoa analysoimalla dataa.....	19
4.2	A/B-testaus.....	19
4.3	Assosiaatiosääntöanalyysi	20
4.4	Kieliprosessointianalyysi	20
4.5	Klusterointianalyysi	20
4.6	Luokitteluanalyysi	21
4.7	Neuroverkkoanalyysi	21
4.8	Regressioanalyysi	21
4.9	Tiedon yhdistäminen ja kombinaatioanalyysi	22
4.10	Tunneanalyysi	22
5	VÄLINEITÄ BIG DATA -TIEDON KÄSITTELYYN	23
5.1	Hadoop.....	23
5.2	Hadoop 1 -ohjelman kehitysversiot analyysi- tai tiedonkäsittelyyn	25
5.3	Apuohjelmistoilla parannettua ohjelmoitavuutta.....	26
5.3.1	Tietohallintaa Hadoop-järjestelmään	26
5.3.2	Tiedonkulkua ja -käsittelyä Hadoopissa parantavat ohjelmat.....	27
5.3.3	Ohjelmia Hadoop-järjestelmän ylläpitoon ja hallintaan	27
5.3.4	Machine learning Hadoop-järjestelmässä	28
5.4	Hadoop 2.0.....	28
5.5	Vastineita avoimen lähdekoodin Hadoop-ohjelmistolle.....	29
6	RAJOITTEET DATAN KÄYTÖLLE	31
6.1	Tiedolle rajoitteita.....	31

6.2	Henkilötietolaki	32
6.3	Tietojen suojaus	33
6.4	Big datan käytön etiikka	33
6.5	Kuinka big dataa on hyödynnetty ostajien etsimiseksi	35
7	POHDINTA.....	36
	LÄHTEET.....	38

ERITYISSANASTO

Big Data	big data, suuren volyymin data
Business Data	liiketoimintatieto, -data
BI	liiketoimintatieto (Business Intelligence)
CRM	asiakkuudenhallintajärjestelmä (Customer Relations Management)
Data	tieto, data
ERP	yrittäjän toiminnanohjaus, - tietojärjestelmä (Enterprise Resource Planning)
ETL	tiedon keräys, muunto ja tallennus (Extract - Transform - Load)
HDFS	hajautettu tiedostojärjestelmä (Hadoop Distributed File System)
Open Data	avoin tieto (lähde)
Open Source	avoin lähdekoodi
RSS	verkkosyöte (blogi, podcast tai uutinen) (Really Simple Syndication)

1 JOHDANTO

Opinnäytetyössä on ollut tavoitteena tutkia big datan hyödyntämistä yrityksen myyntityössä ja markkinoinnissa. Yrityksellä on usein hyödynnettävissä paljon omaa tietoa. Tätä uudelleen käsittelemällä ja yhdistämällä siihen myös muuta kuin yrityksen käytettävissä olevaa tietoa, voidaan saavuttaa uutta osaamista asiakaskunnasta tai oman yrityksen resursseista. Myynnin ennustaminen tuotekohtaisesti on jokaiselle yritykselle hankalaa. Mikäli tietoa asiakkaiden trendeistä on mahdollista saada, on mahdollista varautua paremmin kysyntään ennakolta.

Suomen kielessä sanalla tieto vastataan usein englannin kielen termeihin data, information tai knowledge. Arkikielessä tiedolla viitataan usein juuri dataan, informaatioon ja tietoon. Kielessämme tiedolla on myös arvomerkitys. Tietoyhteiskunta-sanalla on meillä enemmän painoarvoa, kuin informaatioyhteiskunta-sanalla, jota yleisesti käytetään maailmalla samasta aiheesta puhuttaessa. Dataan pitää tiedon ja informaation ohella liittää myös sisältö-, teos- ja tekijänoikeuskäsitteet. (Poikola, Kola & Hintikka 2010.)

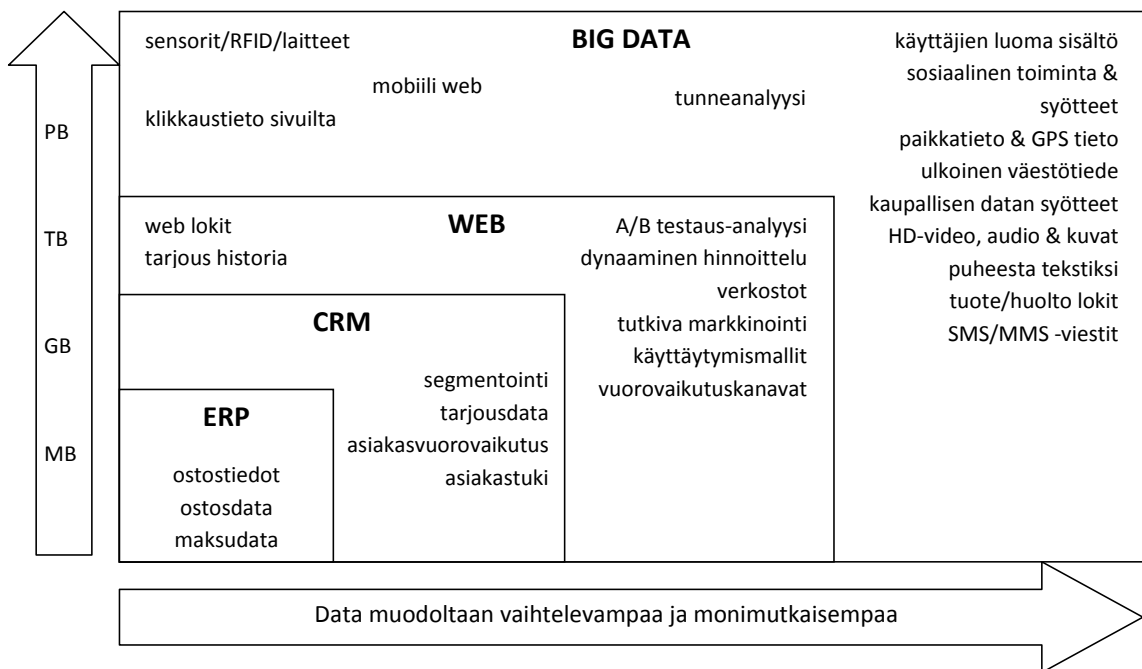
Dataa eli tietoa syntyy tälläkin hetkellä suunnattomia määriä. Sitä tuottavat ihmiset toimiessaan Internetin keskustelupalstoilla tai koneellisesti kulkiessaan erilaisten ilmaisimien ohi. Perinteisesti yritysten data tai tieto, varsinkin jos otetaan tarkasteluun myynti tai markkinointi, koostuu erilaisista transaktioista. Transaktiotiedot kerätään päivän päätteeksi ETL-menetelmillä (Extract, Transform and Load) strukturoituun eli rakenteelliseen tietovarastoon myöhempää tarkastelua varten. Erilaisilla BI-työkaluilla (Business Intelligence) on mahdollisuus tarkastella saatua tietoa ja verrata sitä vaikkapa edellisen vuoden tapahtumiin. Opinnäytetyössä on tarkoitus myös kuvata lyhyesti erilaisia menetelmiä, joilla tätä yrityksen käytettävissä olevaa big dataa voi kerätä ja hyödyntää.

Asiakaskorttien käytöstä kertyy paljon henkilökohtaista tietoa. Tutkin mitä tietoa ja miten näitä tietoja voidaan ottaa tarkastelun kohteeksi. Mukana on myös osio, jossa tarkastellaan suomalaisen lainsäädännön mukaisia mahdollisuuksia hyödyntää ja käsitellä kerättyä tietoa.

2 BIG DATASTA HYÖTYÄ YRITYKSELLE

2.1 Big data

Big data on tietoa, jota ei pystytä prosessoimaan perinteisillä tietokantajärjestelmillä. Tiedostot ovat liian isoja, sisältö muuttuu liian nopeasti tai data on sopimatonta tietokanta-arkkitehtuuriin (Kuvio 1). Tiedon haltuun ottamiseksi on pitänyt valita toisenlaisia tapoja prosessoida sitä. (Big Data Now 2012, 3.) Big data onkin avainasemassa mahdollisuuksineen, kunhan tietoa pystytään keräämään, analysoimaan ja käyttämään liiketoiminnan kehittämiseen (Mohanty, Jagadeesh & Srivatsa, 2013, 8).



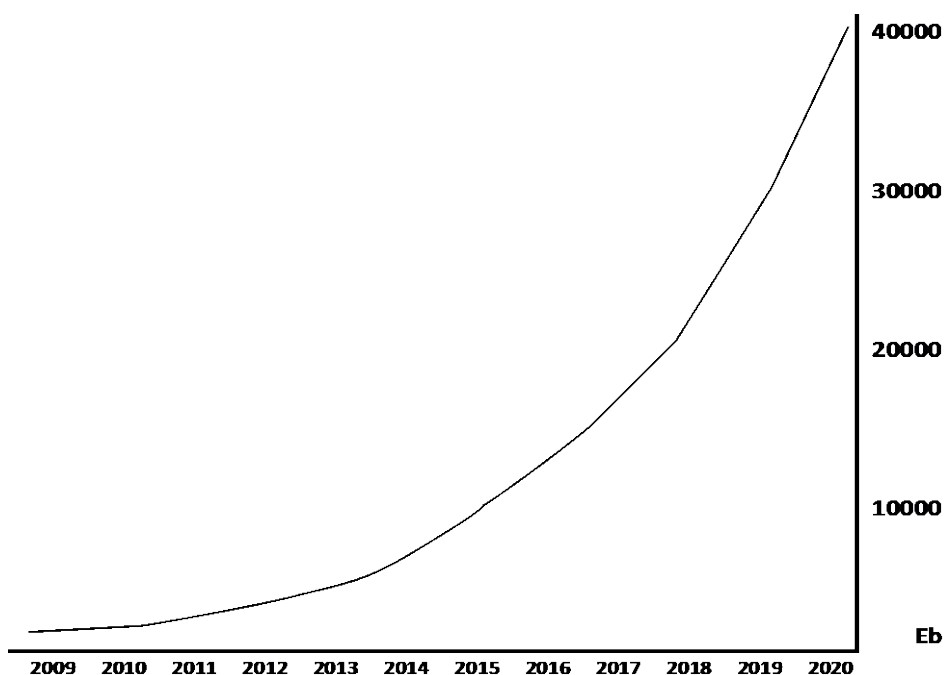
KUVIO 1. Big datan tietomaailma. (Mohanty ym. 2013, 108, mukaisesti.)

Ilmiönä big data on maailmanlaajuinen ja herättää siten paljon epäluuloa. Tietojen keräys epäilyttää ihmisiä ja he pitävät sitä vain heidän yksityiselämänsä vakoiluna. McKinsey'n raportissa (McKinsey 2011, 1,2) kerrotaan kuitenkin big datan eduista kaupankäynnille, valtioille ja niiden asukkaille olevan jo nyt vahvaa näyttöä. Esimerkiksi jos Yhdysvaltain terveydenhoito pystyisi käyttämään luovasti ja tehokkaasti big dataa, se pystyisi tuottamaan 300 miljardin dollarin tuotot, joista 2/3 perustuu 8 % säästöihin terveyskuluissa. Kaupalle on mahdollista kehittää jopa 60 % enemmän liikevaihtoa big

dataa hyödyksi käyttämällä. Euroopan valtiontaloudet pystyisivät säästämään 100 miljardia euroa pelkästään big datan käyttöönottamisella. Laskelmissa ei ole otettu huomioon niitä säästöjä, joita voitaisiin saada petosten estämisellä sekä virheiden ja veroaukkojen tukkimisella. (McKinsey 2011, 1,2.)

Big dataan liitetään usein kolme V-kirjainta. Ensimmäinen V tulee sanasta Volume eli volyymi. Merkitys tälle on johdettu tiedon eksponentiaalisesti kasvavasta määrästä. Toinen V on johdettu Velocity sanasta eli tiedon vauhdista, jolla dataa syötetään sisään ja ulos tietojärjestelmissä. Kolmas V on Variety eli vaihtelevuus, jolla kuvataan datan rakennetta. (Salo 2013, 21.)

IDC:n (International Data Corporation) vuonna 2012 tekemän raportin (The Digital Universe in 2020, 2012) mukaan datamäärän kasvun kerroin maailmassa vuosien 2005 - 2020 välillä on 300 (Kuvio 2). Vuonna 2020 datamäärän arvioidaan saavuttavan 40 000 EB (eksatavu 10^{18}).



KUVIO 2. Arvio datamäärän kasvusta maailmassa vuoteen 2020 mennessä. (The Digital Universe in 2020, 2012, mukaisesti.)

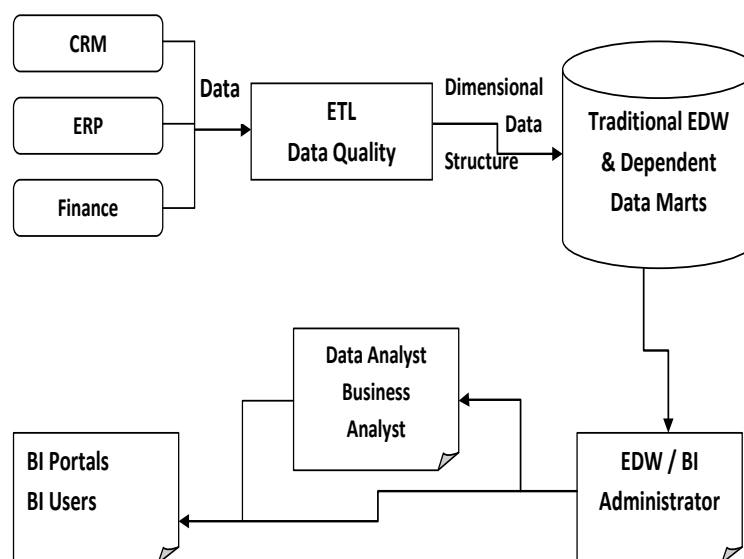
Noin 68 % tämänhetkisestä datasta on kuluttajien tuottamaa tai kuluttamaa tietoa. Muu datamäärä koostuu lähinnä valtioiden, yritysten tai erilaisten anturien tuottamasta datas-

ta. Yksityisten kuluttajien datastakin 80 % kulkee erilaisten yritysten datajärjestelmien läpi. Näille yrityksille jää tuolloin tekijänoikeuksien, yksityisyyden ja säännösten noudattamisen vastuu. Raportissa arvioidaan myös, että vuonna 2020 33 % koko datasta sisältää jotain arvoa, jos se analysoidaan. Yksittäisten henkilöiden itse tuottama datamäärä tulee olemaan paljon pienempi kuin se datamäärä, mitä heistä itsestään tuotetaan. (The Digital Universe in 2020, 2012.)

2.2 Yrityksen oma liiketoimintatieto

Big data on yhdistelmä perustransaktiotietoa ja vuorovaikutteista, erikoistumatonta tietoa. Vuorovaikutteinen, erikoistumaton tieto voi olla esimerkiksi verkkokauppasovelluksen lokitiedostoja, sosiaalisen median klikkauksia ja kommentointia, erilaisista mittausjärjestelmistä tulevaa anturitietoa tai vaikkapa palveluntarjoajien kautta tulevia RSS-syötteitä (Hotti 2012).

Yrityksissä hallitaan yleensä perinteinen transaktiotieto tehokkaasti (Kuvio 3). Datan lähteinä on perinteisesti yritysten asiakas- (CRM), toiminnanohjaus- (ERP) ja rahaliikennetieto (Finance). Joka päivä tiedot tallennetaan tausta-ajona (ETL) yrityksen tietojärjestelmävarastoon (EDW). Tietovarastosta tehdään sitten perinteisillä BI-työkaluilla raportteja ja analyyskejä yrityksen erilaisiin tarpeisiin.



KUVIO 3. Perinteinen transaktiotiedonkulku yrityksessä. (Mohanty ym. 2013, 109, mukaisesti.)

Vaihtelevuudellaan erilaisesta datasta ja tiedostoista koostuva tieto on tullut uudeksi haasteeksi yrityksille. Tämän lisääntyvän tiedon avulla on mahdollista tuottaa huomattavaa lisäarvoa yritykselle. Tiedon käsittely ja hyödyntäminen on mahdollista hoitaa monella eri tavalla. (Mohanty ym. 2013, 1.)

2.3 Kuluttajan tuottama tieto yritykselle

Kauppiaan tietojärjestelmiin syntyy paljon tietoa kuluttajan asioidessa. Jos asiakaskorttina on K-plussakortti tai S-bonuskortti, siitä jää järjestelmiin yksityiskohtaista tietoa asiakkaan ostoista. Kaupat ovat melko varovaisia kertomaan mitä tietoja näistä kuluttajan toiminnoista hyödynnetään. Kauppa on joutunut perustelemaan tietojenlouhintaa asiakaskuntansa suuntaan, kun kyseessä on ollut tuotevirhe ja tuotteiden ostajien etsimiseksi on käytetty transaktio- ja asiakaskorttitietojen yhdistelyä. Yleisesti kaupat kertovat käyttävänsä kerättyä tietoa vain myymäläkohtaisia tuotemyyntiarvioita varten (Talouselämä 31.5.).

Kauppiaas tietää tuotteidensa ostajat. Sosiaalisen median tietojen tai kaupan omien netisivujen lokitietojen avulla kauppiaan olisi mahdollista selvittää kuka ei osta hänen tuotteitaan ja miksi. Tämänkaltaisilla menetelmillä on mahdollisuus parantaa kaupan käyntiä, mahdollistaen myös asiakaskunnasta poimituille pienryhmille suunnattua markkinointia ja tuotevalikoimaa. (Mohanty ym. 2013, 11.)

Suuret verkkokaupat hyödyntävät jo nyt big dataa. Asiakkaiden ostoskäyttäytyminen tai sivun klikkaukset ovat analysoinnin kohteena. Verkkokauppa Amazon on esimerkki dynaamisesta hinnoittelusta, jossa tarjotaan asiakkaille erilaista sisältöä ja hinnoittelua käyttäytymisen perusteella (Mohanty ym. 2013, 15). Samaa menettelytapaa noudattavat myös monet majoitusta tarjoavat verkkosivustot. Vastaavasti halpalentoyhtiöt hinnoittelevat lentolippujen hintoja dynaamisesti, riippuen kysynnästä ja koneen täyttöasteesta.

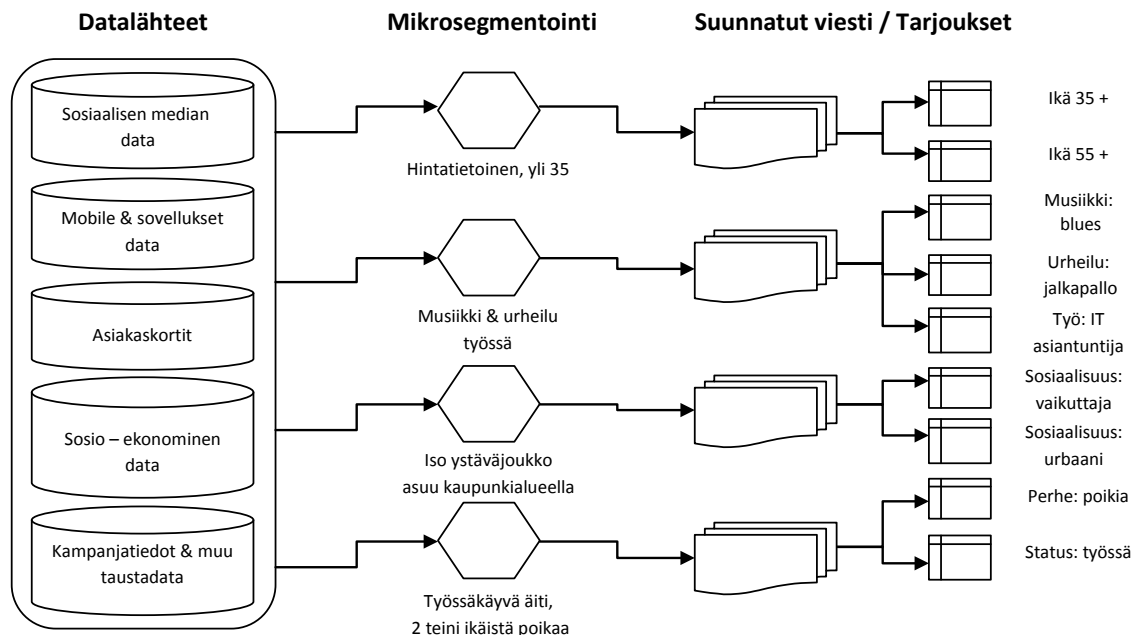
2.4 Datan jatkojalostus uusilla analysoinneilla

Tiedonlouhinta tulee olemaan enenevässä määrin käytössä erilaisten yritysten menestyskamppailun taustalla. Käyttäen saatua tietoa, ne pystyvät suuntaamaan toimensa asi-

akkaiden mielihalujen mukaisesti. Tiedon analysoijia tullaan tulevaisuudessa tarvitsemaan huomattavasti nykyistä enemmän. Syy on erilaistumattoman tiedon tuottaminen maailmassa eksponentiaalisesti, oli sitten lähteenä mikä tahansa tietoa tuottava kohde. (Salo 2013, 21, 144.) Britanniassa on jo hallitustasolla huolestuttu data-analysoijien mahdollisesta vajeesta tulevaisuudessa (ComputerWeekly.com. October 2013). Myös Suomessa vastaavaan koulutusvajeeseen ollaan puuttumassa. Aihe on yhtenä pääkohtana Työ- ja elinkeinoministeriön raportissa 21 polkua Kitkattomaan Suomeen (21 polkua Kitkattomaan Suomeen 2013).

Big data -tallennusmenetelmillä tietoa kerätään talteen, vaikka ei vielä ole tietoa miten sitä analysoidaan. On paljon eri toimijoita, joilla on tarjota työkaluja näiden kerättyjen tietojen tallennukseen ja analysointiin. Toiminnanohjausjärjestelmän, asiakkuudenhallinnan, yrityksen verkkokaupan ja tuotehallinnan tietoja yhdistämällä big data -tietolähteisiin, avautuu yrityksessä uusia mahdollisuuksia analysoida tietoa. Sosiaalisen median eri tietolähteitä yhdistämällä lisäksi, päästään seuraamaan asiakaskuntaa lähes reaaliaikaisesti. Näin on mahdollisuus saada selville, mikä tuote kiinnostaa asiakaskuntaa ja miten esimerkiksi tuotevalikoimaa pitäisi kehittää. (Hotti 2012.)

Big data -analyysimenetelmillä on mahdollisuus kasvattaa myyntiä. Esimerkiksi asiakaskunnan mikrosegmentoinnilla voidaan kohdistaa markkinointi tietyille kohderyhmälle (Kuvio 4). Mikrosegmentointi on ollut tiedossa jo ennenkin, mutta vasta big data on varsinaisesti mahdollistanut tämän hyötykäytön. Saatavilla olevat tiedot ja analyysitekniikka big data -työkaluilla on tehnyt mahdolliseksi jaon esimerkiksi eri mikrosegmentteihin. Vähittäiskauppa saattaisi kutsua tätä myös asiakaskunnan personoinniksi. Kaupalla on mahdollisuus hyödyntää markkinatutkimustietoa, ostoskäyttäytymistä jopa yksittäisten asiakkaiden tasolla, mukaan lukien heidän verkkoklikkaustietonsa. (McKinsey 2011, 68.)



KUVIO 4. Asiakaskunnan mikrosegmentointi big dataa hyödyntäen. (Mohanty ym. 2013, 64, mukaisesti.)

3 MENETELMÄT BIG DATAN HYÖDYNTÄMISEEN

3.1 Mistä tietoa yritykselle

Tietoa kertyy kuluttajakäyttäytymisestä. Transaktiotietoja, mahdollisine asiakaskorttitietoineen, tallennetaan tietovarastoihin. Kaupan tavaravirrasta syntyy logistista tietoa järjestelmiin. Kaikki nämä ovat yksityistä tietoa, jota ei ole kaupan ulkopuolella.

Suomessa henkilötietolaki määrittää melko tarkasti millaista tietoa yksittäisestä kuluttajasta voidaan kerätä. Tiedon siirrolle maan ulkopuolelle on myös omat rajoitteet (Henkilötietolaki, 3§, 4§). Tietojen väärinkäyttö on myös määritelty rangaistavaksi (Rikoslaki, luku 38, 9§). Transaktiotietojen kohdalla pankkitoiminnalle on omat sääntönsä. Esimerkiksi hyvän pankkitoimintasäännösten mukaan velvoitetaan peittämään osin korttitiedot kuiteista, mahdollisten väärinkäytösten estämiseksi.

Julkisena tietona on Suomessakin saatavissa esimerkiksi liikenteen sujuvuustietoja, kaupunkien ilmoitustietoja, tilastotietoja tai ministeriöiden julkaisuja. Aivan julkisia eivät nämäkään tiedot vielä ole, käyttäjän pitää useissa tapauksissa olla kirjautuneena tietojen tilaajana. (Avoin data.)

3.2 Yksityinen eli suljettu tieto yrityksessä

Yrityksellä on usein hyödynnettävissä paljon omaa tietoa. Uudelleen käsittelemällä ja yhdistämällä siihen myös muuta yrityksen käytettävissä olevaa tietoa, voidaan saavuttaa uutta osaamista asiakaskunnasta tai oman yrityksen resursseista (Salo 2013, 32). Aina big data ei ole uutta tietoa, osa siitä on saattanut olla yrityksen tietovarastossa unohtetuna tai hyödyntämättömänä (Mohanty ym. 2013, 14). Yrityksen asiakasrekistereistä on kerättävissä myös paljon sopivaa tietoa jatkojalostusta silmälläpitäen. Yrityksen oma tieto muodostaa myös merkittävän kilpailuedun vastaaviin muihin toimijoihin nähden. Kuluttajien käyttäessä asiakaskortteja oston yhteydessä saadaan taustaa transaktiotapahintaan ja voidaan näin hyödyntää saatua tietoa. Yrityksen verkkosivuston lokitiedot antavat lisätietoa asiakkaiden mielenkiinnosta tuotteita kohtaan. Markkinatutkimusten

käytöllä saadaan suoraan strukturoitua tietoa asiakaskunnan käyttäytymisestä, mikäli otos saadaan riittävän suureksi. (Salo 2013, 33,34.) Tietojen käsittelyssä pitää huomioida henkilötietojen käsittelytarkoitus (Henkilörekisterilaki 8§).

Big data -tietoa on mahdollisuus myös ostaa. Yrityksen ulkopuolinen data on hyödyllistä lisätietoa yrityksen omaan analytiikkaan. Lähteinä voi olla luottokorttiyhtiöiden laskutustiedot, web-operaattoreiden kävijälaskurit ja vaikka puhelinoperaattoreiden paikannustiedot. Usein ulkopuolisen suljetun datan mukana on ostettavissa myös valmista analyysiä käytetystä tiedosta. Näidenkin toimijoiden pitää toki toimia samojen tiedonkäsittelyrajoitusten ja tapojen mukaisesti henkilöiden tunnistamisessa. (Hurwitz, Nugent, Halper & Kaufman 2013, 147.)

3.3 Julkinen avoin tieto eli open data yrityksen saatavilla

Julkishallinnolliset toimijat tuottavat runsaasti erilaista tietoa. Suurimpaan osaan tästä raakatiedoista on aiemmin ollut pääsy vain harvoilla ja jatkojalostukseen on siten saatu vain osa tuotetusta tiedosta. Julkishallinnolla on ollut linjauksena periä korvaus tuottamastaan datasta. Perusteena korvaukselle on vuodelta 1992 oleva maksuperustelaki. (Poikola ym. 2010.)

Julkishallinnon toimialue rajoittaa mahdollisuuksia tiedon avaamiseen julkiseen käyttöön. Tiedon julkaisemisen yhteydessä on tärkeää huolehtia yksityisyydensuojasta, jotta sitä ei tulisi rikottua vapautettaessa tietoa julkiseen käyttöön. (Salo 2013, 36.) Julkishallinnon edustajilla on ollut tapana tuottaa ja käyttää vain omia rekistereitään, koska ne ovat halunneet säästää oman hallinnonalansa kuluista. Tietojen avaaminen julkiseen käyttöön saattaa muuttaa myös eri hallinnonalojen toimintamenetelmiä, kun vastaavia tietoja on saatavissa kustannuksitta toiselta hallinnonalalta. (Poikola ym. 2010.)

Suomessa liikenne- ja viestintäministeriö on mukana avoin data -hankkeessa. Avoin data -hankkeella pyritään avaamaan eri viranomaistahojen tietovarantoja. Sovelluskehittäjille avoin data voisi toimia pohjana uusien sovellusten kehittämiseksi, vaikka mobiilialustalle kansalaisten käyttöön. Ilmatieteen laitos, liikenteen turvallisuusvirasto Trafi ja Viestintävirasto ovat olleet mukana avoin tieto -hankkeessa. Osin syynä on myös julkisuuslain viranomaistehtävä, jonka mukaan viraston on tuotava tieto julkisesti saata-

vaksi, ellei salaamiselle ole erityisiä perusteita. Viranomaistahojen dataa tuotetaan pääosin verovarjoilla, joten avoimella datalla nähdään tuotettavan veronmaksajille uutta hyötyä. Toiminta on vielä alkuasetelmissa ja sovelluskehittäjille ei olla asettamassa esteitä datan käytölle. Avoimen datan myötä eri viranomaistahot pystyvät tekemään paremmin yhteistyötä ja tehostamaan omaa toimintaansa. (Kide 2013, 11–14.)

3.4 Data-analyysi

Big datan analysointiin tarvittavan laitteiston hankinta tai analysointimenetelmien kehittäminen ei ole yksinkertaista. Google tai Yahoo! nettihakuyhtiöinä tietävät, kuinka monimutkaista big datan käsittely ja taltiointi on. Liiketoimintaympäristö tuottaa omat haasteensa. Kurinalaisuutta tarvitaan tiedonkeräilyyn, datan korjaustoimiin, tiedon käytön rajauksiin, tallennukseen ja tiedon edelleen ohjaukseen muille ohjelmistoalustoille ja ohjelmille. Usean maan laeissa vaaditaan henkilötunnusten peittämistä tiedoista ennen luottokorttitransaktioiden siirtoa tiedonnälkäisille tutkijoille. Data-analyysillä petostutkijat etsivät transaktioista merkkejä rahanpesusta tai muusta laittomasta rahaliikenteestä. (Three-Legged Stool.)

Tiedolle pitääkin asettaa rajoitteita, kenellä on oikeus avata ja lukea tiedostoja. Tarvitseeko ylläpitohenkilöstöllä olla oikeus lukea tiedostojen sisältöä? Voiko ohjelmistolla siirtää tiedostoja muualle? Esimerkiksi näille tehtäville saatetaan asettaa rajoituksia. (Olhorst 2013, 69,70.)

Big data vaatii suurta tallennuskapasiteettia ja suurta prosessorimäärää rinnakkaisen tallennus- ja käsittelytoiminnan vuoksi. Alalle on tullut myös paljon toimijoita, jotka tarjoavat big data -tallennus ja -analysointipalvelua pilvipalveluna. Toiminnan yhteydessä mainitaan usein yksityinen pilvi, kun toimintaa suoritetaan yrityksen omilla servereillä ja tallennusmedialla. Julkinen pilvi on määrite silloin kun palvelu ostetaan, joko osin tai kokonaan, ulkoisen toimijan pilvipalvelusta. (Hurwitz ym. 2013, 73.) Tietojen tallentamisessa on huomioitava tallennuskapasiteetin hinta. The Digital Universe in 2020 (The Digital Universe in 2020, 2012) mukaan nykyinen kustannus 1Gb tallennustilaa kohden oli noin 2\$ ja hinta tulee putoamaan siitä viidennekseen tai alle vuoteen

2020 mennessä. Vastaavasti muun laitteiston hintaan arvioidaan muodostuvan kolmannes lisää samana aikajaksona.

Liiketoiminnassa perinteinen raportointi- ja BI-toimintamalli ei pysty taipumaan nopeasti muuttuvaan ja luonteeltaan vaihtelevaan datavirtaan. Big data -analytiikassa on kyse pitkäjänteisestä ja haastavasta tiedon louhintaprosessista, jota joudutaan soveltamaan kulloisiinkin tarpeisiin ja olosuhteisiin. Digital Universe in 2020 -raportissa (The Digital Universe in 2020, 2012) on määritelty, että vain 3 % kaikesta tämän hetkisestä tiedosta on merkitty ja alle puoli prosenttia kaikesta tämän hetkisestä tietomäärästä on analysoitu. Salon (2013, 94) mukaan big datan tulokselliseen analysointiin yrityksissä tarvitaan kolmea eri big data osaaja-aluetta. Ensimmäiseksi tarvitaan liiketoiminta-alueen tunti-joita, jotka tuntevat alan erikoispiirteet. Toiseksi tarvitaan datatieteilijöitä, joilla on hallussaan osaaminen tiedon louhimisesta ja jalostamisesta. Kolmanneksi tarvitaan liikkeenjohdolta näkemystä, sitoutumista, riskinottoa ja pitkäjänteisyyttä tuotteiden ja palvelujen liiketoimintamallin kehittämiseen big datan avulla. (Salo 2013, 94.)

3.5 Tiedosta liiketoimintaa

Big datan ympärille on mahdollisuus kehittyä kokonaan oma liiketoimintaympäristönsä. Mahdollisesti kehittyvät kaupallisesti toimivia yrityksiä, jotka yhdistelevät ja analysoivat toisten yritysten tuottamaa dataa. Erilaisista tuotteista ja palveluista on saatavissa suuria määriä dataa, josta kuluttajien, tuotteiden ostajien ja -toimittajien mieltymyksiä voidaan taltioida ja analysoida. Tämänkaltaisia tietovirtoja voivat tuottaa esimerkiksi:

- kuluttajien ostokset tuotteisiin ja palveluihin
- yritysten globaalit toimitusketjut
- kaupan tuottamat miljoonat transaktiotiedot
- kuluttajille digitaalisia kokemuksia tarjoavat tahot.

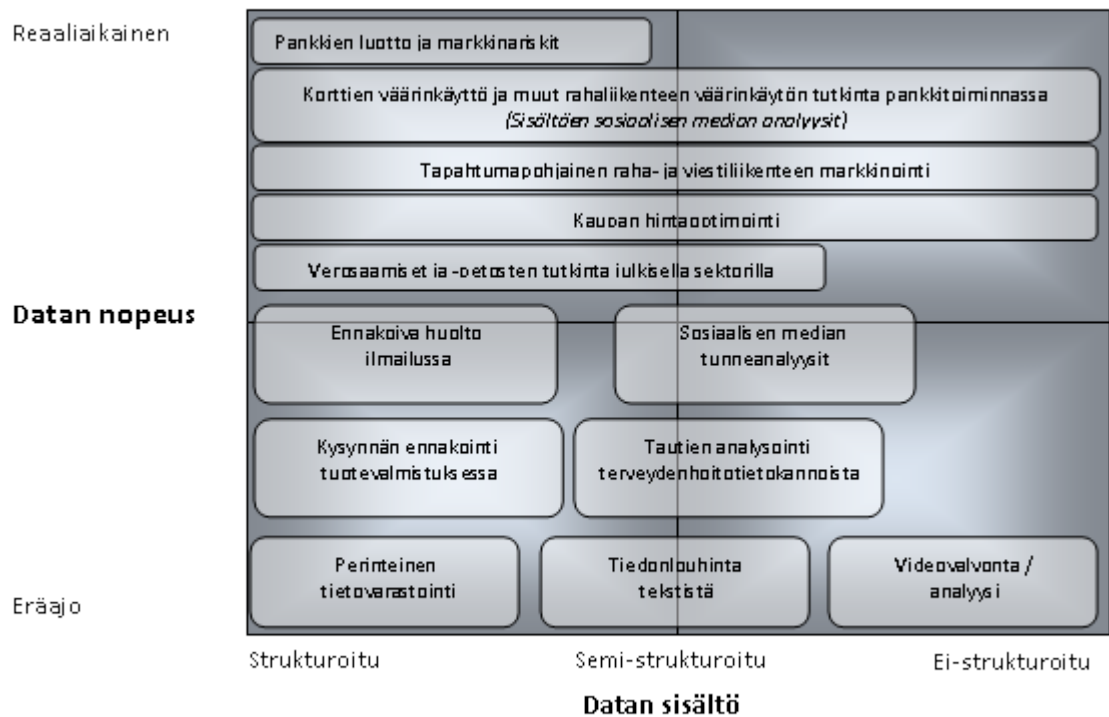
Yritysten oma big data voi olla myös etuna yritykselle. Monilla yrityksillä saattaa olla käytettävissään enemmän omaa big dataa kuin ensiksi ajatellaan olevan. Yrityksillä on pääsy moniin datalähteisiin, joita heidän valmistamansa tuotteet tai palvelut tuottavat. Tietoa voi siirtyä näistä lähteistä joko automaattisesti Internetin välityksellä tai manuaalisesti sitä eteenpäin siirtämällä. Esimerkkinä vaikka autoteollisuuden tuottamien auto-

jen data-anturien tuottama tieto, jota luetaan huoltojen yhteydessä. Datalähteensä ensiksi varmistavilla yrityksillä on tässä mahdollisesti jopa parhaimmat mahdollisuudet taloudelliselle tuotolle. (McKinsey 2011, 6.)

4 BIG DATAN ANALYSOINTIMENETELMIÄ

4.1 Tietoa analysoimalla dataa

Big datan analysointiin käytetään erilaisia tiedonlouhintamenetelmiä. Tiedonlouhintamenetelmillä pyritään erottamaan tietokannoista malleja tai kuvioita tilastollisilla tai koneellisilla menetelmillä (Kuvio 5.). Menetelmiin kuuluu esimerkiksi assosiaatiosääntö-, klusterointi-, luokittelu- ja regressio-analyysi. Tulosten visualisointi on yksi tapa tulkita analyysien tuloksia.



KUVIO 5. Big data analyysit suhteessa datan nopeuteen tai sisältöön (Sas: Roadmaps for the CIO, 9 mukaisesti).

4.2 A/B-testaus

A/B-testaus on vertailuryhmään perustuva tekniikka. Verrataan tutkittavia testiryhmiä, jotta voidaan etsiä niistä eroavaisuuksia. Näin voidaan esimerkiksi määrittää toimenpiteiden vaikutusta tutkimuksen kohteeseen, vaikkapa markkinoinnin vasteisiin. Testaus-tapaa käytetään esimerkiksi verkkokaupan web-sivujen tuloksellisuuden kehittämisessä.

Big data antaa mahdollisuuden suuren testimäärän suorittamiseen, mikäli tiedostojen eroavaisuudet ovat statistiikallisesti riittävät luotettavan lopputuloksen saavuttamiseksi. (McKinsey 2011, 27.)

4.3 Assosiaatiosääntöanalyysi

Assosiaatiosääntöanalyysi on testimenetelmä, jossa tutkitaan muuttujien välisiä suhteita isoissa tietokannoissa. Tyypillinen käyttökohte on ostoskoreista tehtävät analyysit. Kuluttajien ostokset muodostavat tietokannan ja muuttujana on tuotteen ostaminen tai ostamatta jättäminen. Erilaisten algoritmien avulla testataan mahdollisuuksia muuttujien suhteille. Yksi menetelmän käyttötarkoituksista on tiedonlouhinta, jossa kauppias vertailee yhdessä ostettujen tuotteiden säännöllisyyttä. Saatua tietoa käytetään markkinointiin. (McKinsey 2011, 28.)

4.4 Kieliprosessointianalyysi

Kieliprosessointianalyysi on koneoppimiseen (tai keinoälyyn) ja lingvistiikkaan perustuva tekniikka. Analyysitekniikalla tutkitaan luonnollista kieltä. Koneoppimisessa laitteisto hahmottaa sille opetettujen sanojen avulla samankaltaisia sanoja analysoitavasta materiaalista. Koneoppimista ja kieliprosessointia voidaan käyttää apuna monessa eri analysointimenetelmässä. Sosiaalisen median datasta tehdyt tunneanalyysit (sentimental analysis) on yksi käyttökohteista. Tällä tutkitaan kuluttajien reagointia, heihin kohdistettuun markkinointikampanjaan. (McKinsey 2011, 29.)

4.5 Klusterointianalyysi

Klusterointianalyysi perustuu tilastolliseen ryhmäjaotteluun ja näiden tilastollisten ryhmien sisältöä ei etukäteen tunneta. Klusterointianalyysi on tiedonlouhintamenetelmä asiakaskunnan segmentointiin markkinoinnin suuntaamisessa erilaisille kuluttajaryhmille. Se eroaa luokitteluanalyysistä siten, ettei vertailujoukkoa ole annettu ennalta. (McKinsey 2011, 28.)

4.6 Luokitteluanalyysi

Luokitteluanalyysi perustuu saatavan tiedon luokitteluun erilaisiin ryhmiin. Analyysissä ryhmän alkioden sijoittelu ja luokittelu tehdään ennalta annettujen tietojen perusteella. Ryhmittelytietojen ennalta antaminen erottaa luokitteluanalyysin klusterointi-analyysistä. Luokittelu on tiedonlouhintamenetelmä, jonka käyttökohteena on esimerkiksi ostajien käyttäytymissegmentointi. (McKinsey 2011, 28.) Mikrosegmentointi on osa luokittelu-analyysitekniikkaa, mitä on kuvattu aiemmin kuviossa 4.

4.7 Neuroverkkoanalyysi

Neuroverkkoanalyysi perustuu tietokonemallinnukseen, joka löytää tietoja tai malleja datasta. Neuroverkko-laskentamalli on saanut alkuideansa ihmisen hermoverkkoa mallintavasta toiminnasta. Neuroverkko menetelmä soveltuu esimerkiksi ei-lineaaristen mallien tai hahmojen löytämiseen. Menetelmää on mahdollisuus käyttää hahmontunnistukseen tai -optimointiin. Analyysimenetelmällä on mahdollisuus tunnistaa esimerkiksi väärin perustein anottuja vakuutuskorvauksia. (McKinsey 2011, 29.)

4.8 Regressioanalyysi

Regressioanalyysitekniikka perustuu tilastolliseen menetelmään. Menetelmällä analysoidaan muutosta, kun yksi tai useampi lähtöarvoista muutetaan. Analyysimenetelmää käytetään arviointiin tai ennustamiseen. Esimerkiksi myyntimäärän ennustamiseen, kun kaupallinen korkotaso tai markkinaosuus toimii muuttujana. Myös jos tuotteen valmistuksessa tehdään muutoksia, tämän vaikutusta asiakastytyvyyteen voidaan tutkia regressioanalyysillä. (McKinsey 2011, 30.)

4.9 Tiedon yhdistäminen ja kombinaatioanalyysi

Tiedon yhdistämis- ja kombinaatiotekniikka analyysimenetelmänä perustuu usean tietolähteen käyttöön, tehokkaampien ja tarkempien oivallusten saavuttamiseen. Menetelmällä saavutetaan tarkempia tuloksia, kuin jos verrattuna olisi ollut vain yksi tietolähde. Datan käsittelyssä voidaan käyttää signaaliprosessointia tiedon yhdistämiseen. Sosiaalisen median tietoa analysoidaan kieliprosessoinnilla ja tuloksia verrataan reaaliaikaiseen myyntiin. Menetelmällä voidaan tutkia millaisia vaikutuksia markkinointikampanjalla on kuluttajien tuntemuksiin ja ostoskäyttäytymisiin. (McKinsey 2011, 28.)

4.10 Tunneanalyysi

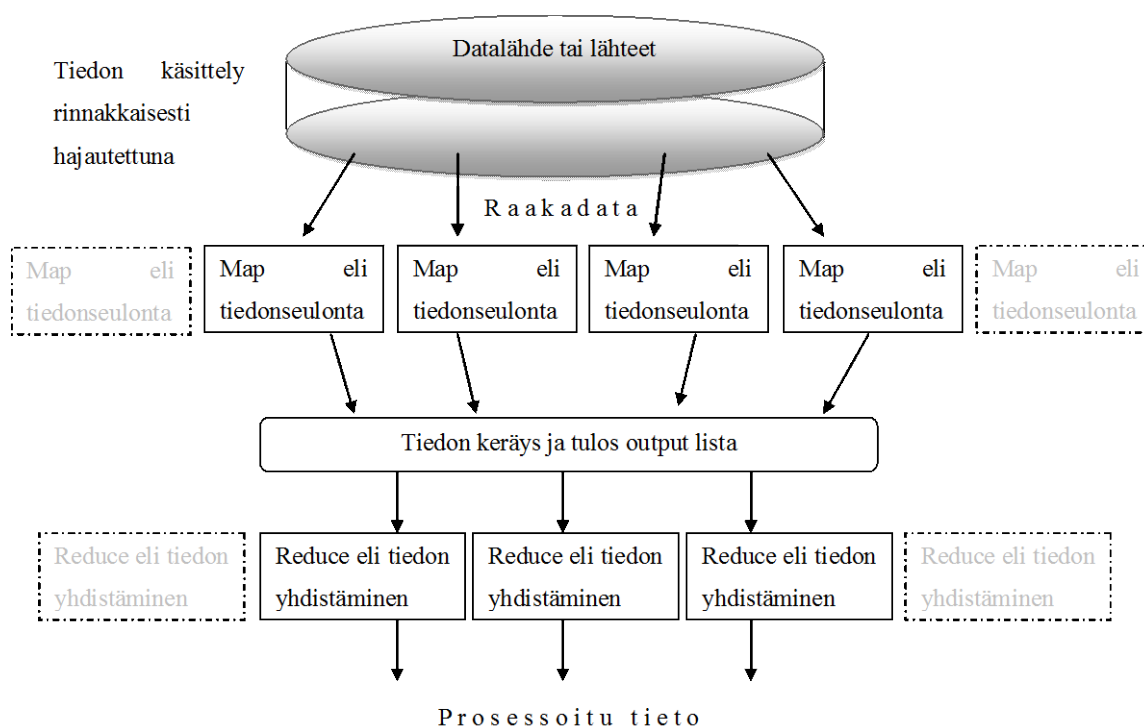
Tunneanalyysimenetelmällä, kieliprosessointia avuksi käyttämällä, saadaan etsittyä ja poimittua subjektiivista tietoa lähdetiedoston tekstimateriaalista. Analyysimenetelmällä tutkitaan tunteen kohdetta, ”polariteettia” (+, - vai neutraali) ja voimakkuutta tutkittavan kohteen suhteen. Esimerkkeinä tunneanalyyseistä ovat sosiaalisen median verkostoihin ja blogeihin, asiakassegmentteihin ja sidosryhmiin kohdistuvat ilmapiirianalyysit, joilla mitataan yrityksen tuotteiden ja palveluiden vastaanottoa. (McKinsey 2011, 30.)

5 VÄLINEITÄ BIG DATA -TIEDON KÄSITTELYYN

5.1 Hadoop

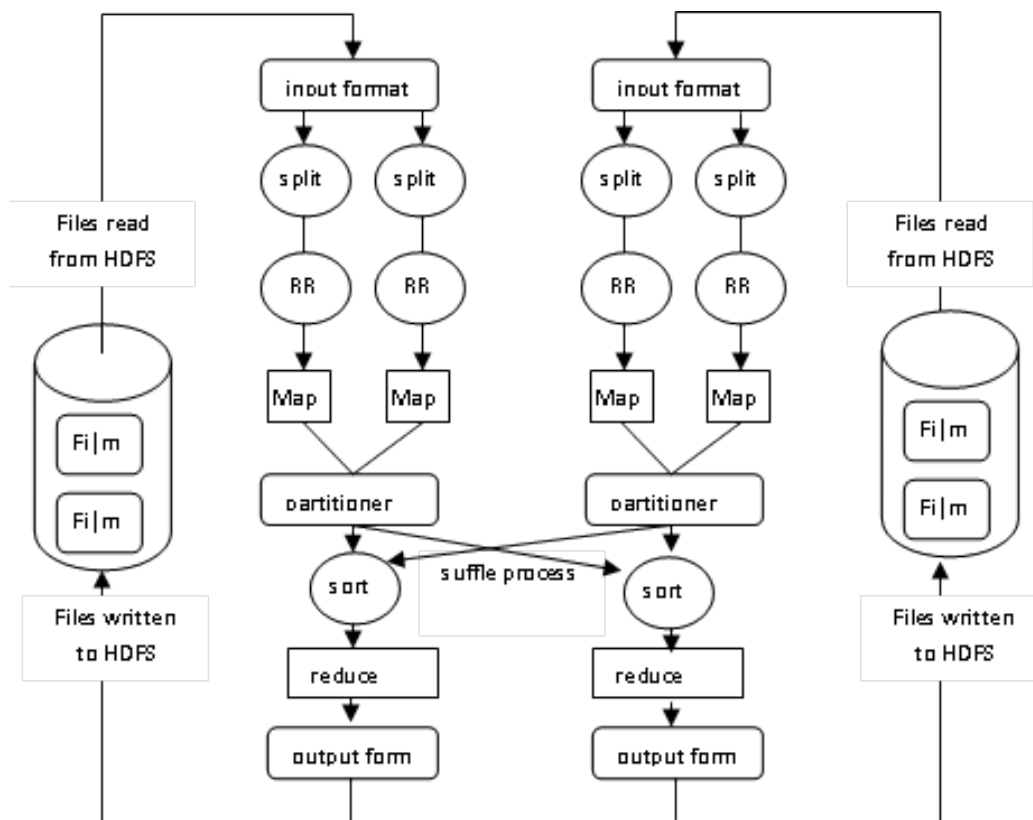
Hadoop-ohjelmistojä voidaan käyttää tekemään analyyssejä, laskentaa tai tiedonlouhimista big datasta. Hadoop-ohjelmistoon perustuvilla ohjelmilla voidaan hyödyntää sekä strukturoitua että ei-strukturoituja tietoja. Usein hakua kuvataan termeillä SQL- (Structured Query Language) tai NoSQL-haku (Not only SQL).

Hadoop on Apache-ohjelmistoprojektissa (Apache Software Foundation) ylläpidetty avoimen lähdekoodin ohjelma. Alun perin ohjelma on Yahoo!-hakukoneyhtiössä toimineen Doug Cuttingin kehittämä ohjelma suurten tietomassojen helpompaan tutkimiseen. (Hurwitz ym. 2013, 111, 112.) Keskeinen toiminnallisuus on tietojen tallennus ja käsittely HDFS (Hadoop Distributed File System) (Kuvio 6.), jota käytetään tiedon hajauttamiseen ja ylläpitoon isoina 64MB tiedostoina. HDFS on Googlen alun perin kehittämä järjestelmä suurten tiedostojen käsittelyyn. (Warden 2011, 9, 18.)



KUVIO 6. Hadoop Map- ja Reduce- perusohjelmien toimintokuva. (Hurwitz ym. 2013, 106, mukaisesti.)

Toinen yleinen ohjelma on MapReduce, jota käytetään tietojen louhintaan. Haut toteutetaan useissa rinnakkaisissa HDFS-klustereissa, käyttäen Java-ohjelmointikieltä. (Kuvio 7.)



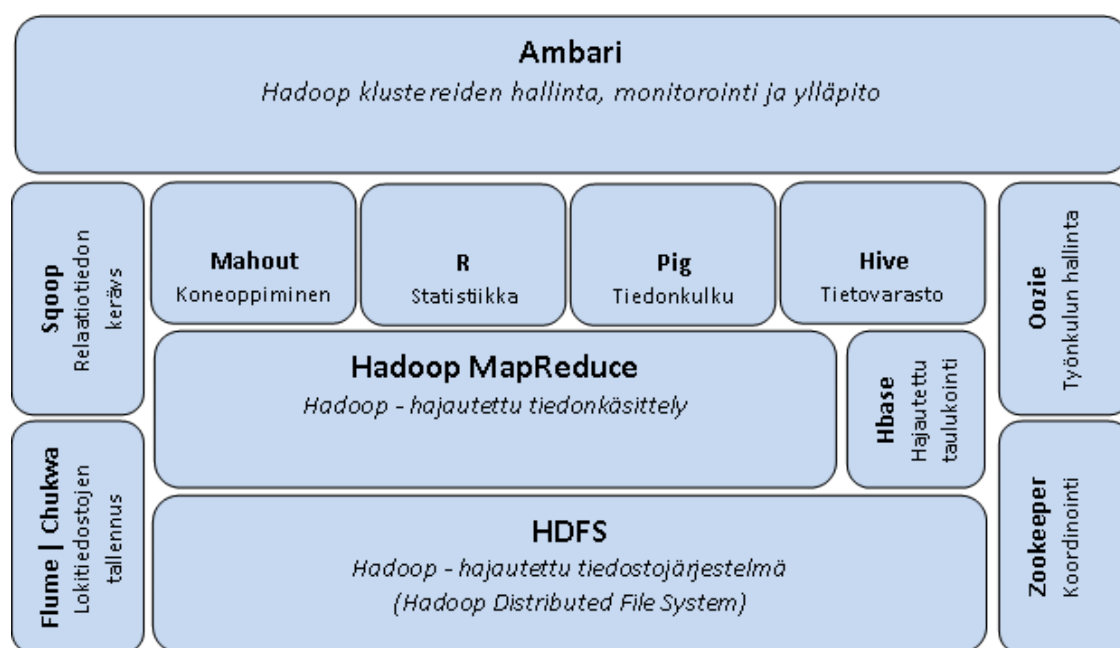
KUVIO 7. MapReduce-ohjelman toimintokuvaus (kuvassa kaksi solmua klusterista).
(Hurwitz ym. 2013, 117 mukaisesti.)

Toiminnan kulku lähtee käyttäjän tekemästä hausta MapReduce-ohjelmalla. Ohjelmakulku jatkuu, kunnes tulokset on kirjoitettu takaisin HDFS-tiedostoon. HDFS-tiedosto jaetaan (split) kaikkien solmujen kesken. Tämän jälkeen suoritetaan solmuissa haluttu haku (map), joista tuloksena on avain-arvo pari. Avain ja arvo -tulokset lajitellaan (shuffle) solmujen kesken. Tulokset järjestellään (sort), yhdistetään (reduce) ja kirjoitetaan takaisin HDFS-tiedostoon. Samanaikaisesti voidaan tehdä hakuja tallennetusta tiedosta useissa solmuissa. (Hurwitz ym. 2013, 22, 117–119.) Toisin kuin normaali tietokantatalennus ja -luku, HDFS ei tee indeksointia tallennuksesta. Yksinkertainen random access -tyyppinen tiedostonhaku ei HDFS -tiedostossa onnistu. (Mohanty ym. 2013, 41.)

Hadoop-ohjelmistosta ei sellaisenaan ole varsinaista ratkaisua yrityksen tietolähteeksi. Se on monimutkainen ja MapReduce-prosessissa suoritettavat haut vaativat paljon ohjelmakoodin kirjoittamista. Vasta näiden toimenpiteiden jälkeen on mahdollisuuksia hyödyntää sen tuloksia raportointiin tai analyysiin. (Hotti 2013.)

5.2 Hadoop 1 -ohjelman kehitysversiot analyysi- tai tiedonkäsittelyyn

Hadoop-ohjelmiston varaan on kehitetty useita erilaisia kaupallisia jatkosovelluksia. Näissä kehitysversioissa on yksinkertaistettu tietojen hallintaa ja tallennusta, joilla on myös päästy parantamaan tietojen käsittelyn laatua (Kuvio 8.). Ohjelmilla on saavutettu nopeampia hakutuloksia. Analyysien tekijöille on saatu helpompia tapoja suorittaa tietohakuja, käyttämällä erilaisia ohjelmointikieliä Java-kielellä ohjelmoinnin sijaan.



KUVIO 8. Hadoop-perusohjelmiston ympärille on rakennettu toiminnallisuutta parantavia ohjelmistoja (Getting Started with Hadoop Planning Guide, 8 mukaisesti).

5.3 Apuohjelmistoilla parannettua ohjelmoitavuutta

Pig eli Pig Latin, on Yahoo!':n kehittämä, helposti omaksuttava Hadoop-ohjelmointikieli. Pig-ohjelmointikieli on yksi Apache-projektin avoimen lähdekoodin ohjelmistoista Hadoopille. Pig-kielellä on mahdollisuus tehdä monimutkaisia haku-kyselyjä Hadoop-tietovarastosta, suorittaen ne halutussa järjestyksessä. Pig-ohjelmointikieltä on verrattu usein big datan ”ilmastointiteipiksi”, ohjelmointiin käytet-
tävän skriptikielen helppouden vuoksi. (Warden 2011, 13.)

Hive on Facebookin kehittämä Hadoop-pohjainen tietovarastojärjestelmä. Ohjelma mahdollistaa SQL-tyyppiset kyselyt, jotka muunnetaan rinnakkain hajautetuiksi Map-Reduce-ohjelmalla tehtäviksi. Esimerkiksi SQL-osaajat voivat tehdä BI-kyselyjä tietovarastoon ilman MapReducen osaamista. (Hurwitz ym. 2013, 118.) Vaikkakin kyselyt tehdään SQL-menetelmillä, yksinkertaiset haut voivat kestää minuutteja Hadoop-tiedostojen suuren koon vuoksi (Warden 2011, 12.).

Cascading-ohjelma on tarkoitettu suorittamaan monimutkaisia työnkulkuja hakujen yhteydessä. Java-rajapinnan avulla laaditaan graafinen muoto halutusta työnkulusta ja ohjelmisto toteuttaa tarkistusten jälkeen tämän toiminnon Hadoop-klusterissa. Mrjob on vastaava kevyempi ohjelmistokehysversio käyttäen Python-ohjelmointikieltä. (Warden 2011, 13.)

5.3.1 Tietohallintaa Hadoop-järjestelmään

HBase on kehitetty avoimen lähdekoodin projektiksi Googlen toimesta. HBase pohjautuu HDFS-tiedostomalliin. HBasen avulla Hadoop-järjestelmästä voidaan lukea ja kirjoittaa tietoja, vaikka MapReduce-ohjelma toimii siinä taustalla. HBase on siten ei-relaatiomuotoisen tietokannan nopea hakutyökalu Hadoopille. Sen avulla voidaan lisätä kaupallisia toimintoja Hadoop-järjestelmään, koska se mahdollistaa usealle käyttäjälle tiedon päivittämisen, lisäämisen ja poistamisen, ilman järjestelmän pysäyttämistä. (Warden 2011, 9.)

Sqoop on yhdistämisohjelmisto tiedon siirtämiseksi ei-Hadoop-pohjaisista tietokannoista, kuten relaatiotietokannoista tai tietovarastoista, Hadoop-tietokannan käytettäväksi. Se antaa mahdollisuuden käyttäjälle määrittää tiedon sijainnin Hadoopin sisällä. Sqoop mahdollistaa relaatiomallisten tietojen siirron takaisin käyttäjän määrittämään sijaintiin, esimerkiksi Oraclen, Teradatan tai muun vastaavan relaatiotietokantaan. (Hurwitz ym. 2013, 126.)

Flume on ohjausohjelmisto, jolla ohjataan tietoa kaikista yrityksen IT-järjestelmistä Hadoopin käytettäväksi. Lähteenä voi olla esimerkiksi web-servereitä, ohjelmistoservereitä, mobiilijärjestelmiä. (Hurwitz ym. 2013, 119.)

5.3.2 Tiedonkulkua ja -käsittelyä Hadoopissa parantavat ohjelmat

Avro, tarkemmin Apache Avro, on ohjelmisto tiedon jäsentelyyn ja muuntamiseen sarjamuotoiseksi. Tietoihin sisällytetään mukaan malli tiedon rakenteesta ja sen käyttöliityntätiedot. Näin on parannettu mahdollisuuksia käsitellä satunnaisia tietomuotoja tietohakujen yhteydessä. (Warden 2011, 42.)

Oozie-ohjelma mahdollistaa työnkulun määrittelyn prosessissa. Käyttäjän voi tehdä eri ohjelmistokielillä hakuja ja Oozie-ohjelmalla ne ohjataan järjestykseen. Uusille hauille, vaikka MapReduce-ohjelmaa käyttäen, voi asettaa toteutusehdot, jotka perustuvat edellisten hakujen valmistuneisiin tuloksiin. (Hurwitz ym. 2013, 119.)

Zookeeper-ohjelmaa käytetään koordinoimaan ja suorittamaan nimeämispalveluja Hadoopin klustereille. Koska Hadoop-klustereiden pitää synkronoitua toistensa kanssa, pitää järjestelmän tietää, miten se käyttää niitä ja miten ne ovat konfiguroituja. (Big Data Now. 2012, 14.)

5.3.3 Ohjelmia Hadoop-järjestelmän ylläpitoon ja hallintaan

Ambari-ohjelmalla on Hadoopin ytimeen tuotu hallinta- ja valvontaominaisuuksia. Ylläpitäjillä on ohjelmaa käyttäen mahdollisuus päivittää Hadoop-klustereita, määrittää ja

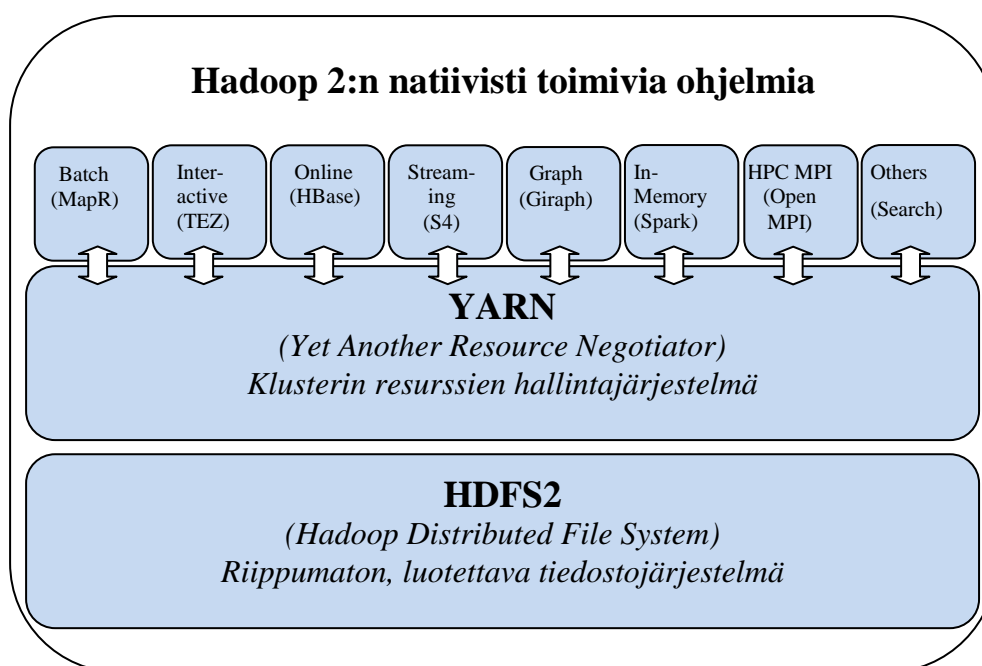
monitoroida koko Hadoop-järjestelmää. API -ohjelmarajapinnan kautta se on myös mahdollista integroida muihin järjestelmien hallintatyökaluihin. (Big Data Now 2012, 14)

5.3.4 Machine learning Hadoop-järjestelmässä

Mahout on tiedonlouhintaohjelmisto Hadoop-järjestelmään. Tiedonlouhinta on suurista tietomääristä asioiden erittelyä tai etsintää. Mahout-ohjelma voi käyttää yleisimpiä tiedonlouhinta-algoritmeja, eli tarkkaan määriteltyjä käskyjä rinnakkain verkotetussa Hadoop-ympäristössä. Testaamisessa Mahout-ohjelman avulla voidaan Hadoop-järjestelmässä suorittaa esimerkiksi yhtäaikaaisesti usealla solmulla regressiotestausta, jolla pyritään löytämään ohjelmistomuutosten yhteydessä tehtyjä virheitä. Hadoop-järjestelmää voi käyttää Mahout-ohjelmiston yhteydessä myös matemaattisen mallinnuksen työkaluna. Verkotettu rinnakkainen ympäristö mahdollistaa samaan aikaan usean yhtäaikaisen suorituksen. (Warden 2011, 31; Hurwitz ym. 2013, 119.)

5.4 Hadoop 2.0

Hadoopin uudemmassa kehitysversiossa Hadoop 2.0 on pyritty kehittämään rinnakkaisesta toiminnallisuutta paremmaksi, irti Hadoop perusversion MapReduce -ohjelman tiedostohallinnasta. Hadoop 2.0 tiedosto klustereiden hallintaan on uusi järjestelmä YARN (Yet Another Resource Negotiator). YARN avulla saadaan erotettua töiden hallinta ja ajoitukset omaksi kerroksekseen prosessointikerroksen alle, jolloin Hadoop voi suorittaa eri sovelluksia (Kuvio 9.). Muutos mahdollistaa Hadoop 2 -järjestelmälle laajemman skaalan analyysi- tai yrityssovellusten käytössä. Esimerkiksi tapahtumaprosessoinnin, tiedostojen streamaus, reaaliaikaohjelmistojen ja toimintasovellusten ajo tulee mahdolliseksi Hadoop 2 -järjestelmässä yhtä aikaisesti. (TechTarget: Hadoop -2.)



KUVIO 9. Hadoop 2 version prosessikerrokset. (Hadoop Version 2: One Step Closer to the Big Data Goal mukaisesti.)

Hadoop 2:ssa on tiedostojen hallinnassa päästy eroon yhden niminoodin käytöstä. HDFS-tiedostojen high-availability -toiminnon avulla annetaan käyttäjille mahdollisuus riippumattomien niminoodien käyttöön ja samalla mahdollistetaan useiden sovellusten yhtäaikaista suoritus. Tällä toiminnallisuuden kehittämisellä on myös estetty yhden niminoodin kaatumisen aiheuttama toiminnallisuuskatko (Single Point of Failure). Hadoop 2:ssa toiminnallisuutta on parannettu esimerkiksi Microsoft Windows -järjestelmän suuntaan ja read only -varmuuskopioiden teko snapshot-toiminnallisuudella on tuotu myös mukaan Hadoop-järjestelmään. Hadoop 2 -kehitysversiona on pystytty kuitenkin pitämään binäärinen yhteensopivuus Hadoop 1 -versioiden suuntaan. (TechTarget: Hadoop -2.)

5.5 Vastineita avoimen lähdekoodin Hadoop-ohjelmistolle

InfoSphere Streams on IBM yhtiön sovellus big datan käsittelyyn. Sen avulla voi suorittaa reaaliaikaista tai ennustavaa analyysiä erityyppisille tiedoille. (Hurwitz ym. 2013, 197.)

MapR-ohjelmisto, kaupallinen yritysversio Hadoop-ohjelmistosta. Siinä on kehitetty oma tiedostojärjestelmä HDFS tilalle. Muutoksilla on pyritty parantamaan käytettävyyttä ja yksinkertaistamaan tiedostojen siirtoa. (Warden 2011, 14.)

S4 on Yahoo!':n 2010 kehittämä vastine Hadoop-järjestelmälle. Koko alustan ohjelmointi on tehty Java-kielellä. Järjestelmään tapahtumia lähettävät ja vastaanottavat käyttäjät voivat suorittaa sen avulla komentoja millä tahansa ohjelmointikielellä. S4 on suunniteltu todella suuriin kokonaisuuksiin ja suoritustehoa voi lisätä lineaarisesti lisäämällä datasolmuja klustereihin. (Hurwitz ym. 2013, 197.)

Storm on Twitterin käyttämä sovellus datavirtojen tutkimiseen. Kuten S4 sitä voidaan käyttää eri ohjelmointikielillä. Storm on tarkoitettu reaaliaikaiseen analysointiin, jatkuvan laskentaan tai integrointiin. Twitterin lisäksi on muita kaupallisia toimijoita, jotka käyttävät sovellusta big datan käsittelyyn. (Hurwitz ym. 2013, 197.)

6 RAJOITTEET DATAN KÄYTÖLLE

6.1 Tiedolle rajoitteita

Henkilötietolaki määrittää henkilörekisteriä ja sen sisältöä. Yrityksissä siihen luetaan eri asiakassuhteiden hoitoon muodostetut asiakasrekisterit. Kunnalle säädetyissä tehtävissä ja valtion lakisääteisissä tehtävissä muodostuu erilaisia henkilörekistereitä, jotka kuuluvat saman lain valvonnan piiriin. Henkilörekisterin ylläpitäjän pitää myös huolehtia, ettei esimerkiksi henkilötunnusta käytetä tarpeettomasti. (Tietosuoja.fi \ 1) Rikoslain 38 luvun 9§:ssä on säännös tahallisesta henkilörekisteririkoksesta, mikäli on käsitellyt henkilötietolain vastaisesti henkilötietoja, antanut rekisteröidylle väärää tietoa tai siirtänyt henkilötietoja Euroopan unionin tai talousalueen ulkopuolelle. (Rikoslaki, luku 38, 9§.)

Big datan hyödyntämisessä onkin otettava selkeästi huomioon tietojen käyttötarkoitus, tietosisältö ja tietovirrat. Henkilötietolain mukaisesti on tehtävä toiminta-analyysi, jossa määritellään seuraavia asioita:

- Henkilörekisterin käyttötarkoitus
- Tietosisältö ja -rakenne
- Tietolähteet
- Tiedon käyttö ja luovutukset
- Tiedon säilyttäminen
- Kuinka rekisteröityjen oikeudet toteutuvat.

Mikäli tietoja siirretään ulkopuoliselle taholle, pitää nämä samat seikat huomioida toimeksiantosopimuksessa. Tietoturvallisuuden varmistamisen velvoite kuuluu tietysti kaikkiin toteutuksen ja tiedon käsittelyvaiheisiin. (Tietosuoja.fi \ 2.)

Big data -lähteisiin luettavalle kameravalvonnalle on laissa myös omat tarkat sääntönsä. Esimerkiksi kuvaamalla saatujen tallenteiden käyttö ja muu tarkoitus pitää suunnitella ja toteuttaa ottaen huomioon henkilötietolain säädökset. Laissa on myös määritelmä tallenteiden hävittämisestä viimeistään vuoden kuluttua tallennuksesta, ellei sille ole muuta erityistä syytä. (Tietosuoja.fi \ 3.)

6.2 Henkilötietolaki

Suomessa henkilötietojen käsittelyä rajaa henkilötietolaki (523/1999). Laki on alkupe-
räisessä muodossaan tullut voimaan 1.12.2000. Laki perustuu Euroopan Unionin henki-
lötiedodirektiiviin 95/46/EC. Lain tarkoituksena on suojata ja turvata yksityiselämää
sekä henkilön perusoikeuksia. Lailla pyritään myös edistämään hyvän tietojenkäsittely-
tavan kehittämistä ja noudattamista. Lakia sovelletaan henkilötietojen automaattiseen
käsittelyyn tai mikäli henkilötiedot muodostavat osan rekisteriä. (Henkilötietolaki, 1§,
2§.)

Lain määritelmissä on henkilötietojen keräämiseen ja käyttötarkoituksiin liittyviä mää-
reitä, jotka voivat liittyä tietojen käsittelyyn tai tallentamiseen. Lain soveltaminen alkaa
silloin kuin rekisterinpitäjän toimipaikka on Suomen alueella tai muutoin Suomen oi-
keudenkäytön piirissä. Lakia sovelletaan myös, jos rekisterinpitäjällä on henkilötietoja
käsitteleviä laitteita Suomessa, vaikka rekisterinpitäjällä ei ole toimipaikkaa Euroopan
unionin jäsenvaltioiden alueella. (Henkilötietolaki, 3§, 4§.)

Laissa on erilliset määritteet henkilötietojen käsittelyn suunnitteluun ja käyttötarkoituk-
seen. Määritteissä on yksitoista lukua erillisine pykälineen, joissa on muun muassa
suostumus henkilötietojen käsittelyyn, lakisääteiset velvoitteet henkilötietojen käsitte-
lylle, tiedon tarkoituksenmukaisuus työnantajalle tai maksupalveluun. Pykälässä kah-
deksan, määritteessä viisi otetaan kantaa asiakkuussuhteeseen ja tällä pitää olla asialli-
nen yhteys rekisterinpitäjän toimintaan. Määritteissä on otettu kantaa myös tietojen luo-
vuttamiseen edelleen. Lain kymmenennessä pykälässä otetaan myös kantaa rekisterin
ylläpitäjään, tarkoitukseen, luovutetaanko tietoja edelleen ja kuinka tiedot aiotaan suoja-
ta. (Henkilötietolaki, 6§-10§.)

Henkilötietolaissa on myös vielä useita kohtia, jotka koskevat henkilötunnuksen käsitte-
lyä. Laki ottaa myös kantaa henkilötietojen siirtoon Euroopan unionin tai talousalueen
ulkopuolelle. Henkilötietojen siirto on mahdollista, jos maassa on riittävä taso tie-
tosuojalle. (Henkilötietolaki, 13§, 22-23§.)

6.3 Tietojen suojaus

Big datan tuottaman tiedon suojaus on yritykselle tärkeää. Ehdotonta se on esimerkiksi aloilla, joissa käsitellään terveystietoja ja henkilötietoja yhdessä. Tiedon hallinnoinnissa pitää täyttää samanlaiset vaatimusedellytykset kuin muunkin yksityistiedon käsittelyssä on ja erityisesti henkilötiedot pitää suojata. Dataan käsiksi pääsyn edellytykset pitää määritellä sovellus- ja käyttäjärooolitasolla. Kriittisen tiedon salaaminen on yksi mahdollinen este tiedon väärinkäytölle. Mikäli data salataan, ei salausavaimia saa säilyttää samoilla palvelimilla tiedon kanssa. Turvallisuusmääritteet on huomioitava koko ketjulle jo alusta alkaen, eikä vasta jälkijunassa. (Hurwitz ym. 2013, 19, 52.)

The Digital Universe in 2020 (The Digital Universe in 2020, 2012) raportissa mainitaan suojatun datan osuuden olleen noin kolmannes kaikesta datasta vuonna 2010. Vuoteen 2020 suojaamista vaativan datan osuus kasvaa noin 40 % kaikesta tiedosta. Raportin mukaan maailmanlaajuisesti vain noin puolella suojausta vaativalla tiedolla on suojausta tällä hetkellä. (The Digital Universe in 2020, 2012.)

6.4 Big datan käytön etiikka

Big dataa tutkittaessa on otettava huomioon neljä seikkaa tietojen käsittelyssä, koskien sekä yksityishenkilöitä tai organisaatioita. Ensimmäisenä on huomioitava onko tiedon lähde yksityinen vai julkinen. Toiseksi on otettava huomioon yksityisyys eli kenellä on oikeus tietoon. Kolmanneksi on huomioitava tiedon omistajuus. Tiedon käsittelijän oikeudet siirtää tietoa eteenpäin tai tiedon käsittelijän dataa koskevat rajoitteet. Neljänneksi tulee huomioitavaksi kaiken kerätyn tiedon luoma maine ja onko kerääntyneeseen tietoon perustuva arviointitieto edes luotettavaa. (Davis & Patterson 2012, 2-3.)

Henkilön terveydellisellä tai taloudellisella tilalla on eniten merkitystä, kun arvioidaan lääketieteellistä hoitoa tai sopivan luoton tarjoamista. Kuitenkin tavallisen kuluttajan kannalta nämä kaksi ovat kaikkein herkimmät alueet. Vastaavasti tietoturva nousee esille eli kuinka suojata tämäntyyppinen arka tieto ja pitää se yksityisenä. (McKinsey 2011,11.)

Sosiaalinen media antaa mahdollisuuden liittyä erilaisiin yhteisöihin, joiden osallistujat ovat ympäri maailmaa. Facebook on ehkä tunnetuin esimerkki, mutta on muitakin yhteisöjä, jotka keräävät jäsenistään yksityiskohtaista tietoa. Käyttäjän syöttäessä päivitystä sivulle, hän ei ajattele tiedon omistukseen liittyviä seikkoja. Järjestelmien tallentaessa kaiken, hän ei ajattele syötetyn tiedon vaikutusta näkemykseen meistä tai sen vaikutuksesta maineeseemme tulevaisuudessa. Henkilön itse säännöllisesti tallentama tieto käyttäytymisestään ja ajatuksistaan antaa mahdollisuuden näiden tietojen hyödyntämiseen tai väärinkäyttöön monellekin taholle. Kaupallisessa mielessä tietojen käytön hyödyt on jo huomattu, mutta mitä tapahtuu tulevaisuudessa kun yksityistietoja louhitaan, yhdistellään, myydään, tai uudelleen myydään ja yhdistellään keskenään (Davis & Patterson 2012, 7.).

Aina kun suurempia määriä dataa liikkuu organisaatioiden rajojen yli, pitäisi mukaan liittää asiat tiedon yksityisyydestä, turvaamisesta, yksityisestä omistamisesta ja luotettavuudesta. Kuluttajien osalta yksityisyys on tekijä, jonka arvo tulee nousemaan big datan yleistessä. (McKinsey 2011, 63.)

Big datan yhteydessä nouseva tiedon taloudellinen merkitys herättää joukon oikeudellisia kysymyksiä, yhdistettynä sen eroavuuteen muusta varallisuudesta. Tietoa voidaan kopioida täydellisesti ja sitä on helppo yhdistää muun tiedon yhteyteen. Samaa tietoa voidaan käyttää yhtäaikaaisesti monessa eri paikassa. Kaikki nämä erottavat tiedon perinteiseen varallisuuteen verrattuna. Pitäisikö kysyä, kuka omistaa yksittäisen tiedon ja siihen kuuluvat oikeudet. Miten voi selvittää tiedon oikeudenmukaisen käytön? Kuka on vastuussa, jos epätarkkoja tai väriä tietoja yhdistämällä päädytään väriin tuloksiin? Tämän kaltaisten kysymysten oikeudelliset vastaukset vaativat selvitystä ennen big datan täydellistä hyödyntämistä. (McKinsey 2011, 84,95.)

Viimeaikaiset esimerkit tietovuodoista osoittavat, että myös yksityiset tiedot voivat joutua tietovuotojen kohteeksi siinä kuin valtiolliset salaisuudetkin. Näiden vakavien tietovuotojen valossa on tietoturvaan kiinnitettävä yhä enemmän huomiota (McKinsey 2011, 11.). Aihe oli erittäin ajankohtainen vuonna 2013. Uutisissakin oli tuolloin noussut esille vaikkapa Yhdysvaltojen tekemä vakoilu viestiliikenteeseen tai sosiaalisen median palveluihin. Näissä kaikissa on käytetty menetelminä tiedon louhintaa, jopa suoraan saapuvasta datavirrasta.

6.5 Kuinka big dataa on hyödynnetty ostajien etsimiseksi

Tiedossa on ainakin kaksi tapausta, joissa kauppaketjut ovat käyttäneet asiakastietoaan jäljittääkseen tiettyjen tuotteidensa ostajia. Näissä kummassakin tapauksessa on pyritty estämään ostajille aiheutuvia terveyteen kohdistuvia seurauksia tuotevirheen vuoksi.

S-ryhmä käytti keväällä 2013 asiakasomistajarekisteriään selvittääkseen pakastevihannesten ostajia, yhdistämällä kuittirekisterin tiedot ja asiakasrekisterin henkilötiedot. S-ryhmä käyttää normaalisti kuittitietoja rekisteröityneille asiakkailleen maksettaviin bonuksiin ja maksutapaetuihin. (Talouselämä 23.5.2013.) Vastaavalla tavalla on aiemmin toiminut K-ryhmä, kun sen tuotteessa oli vastaavan tyyppinen sisällöllinen virhe (Yle Kotimaa 30.10.2011.).

Ymmärrettävästi kaupanalan toimijat eivät suuresti kerro, kuinka paljon he tietävät kulluttaja-asiakkaan tekemisistä. Esimerkiksi S-ryhmä toteaa rekisteriselosteessaan vuonna 2013 keräävänsä tietoa lähinnä valikoiman seurantaan ja kehittämiseen tuote- ja tuoterhyhmätasolla, sekä tietoihin S-ryhmän palvelujen käytöstä (S-rekisteriseloste.).

7 POHDINTA

Big datan avulla on mahdollisuus tuottaa taloudellisia hyötyjä yritykselle. Jotta taloudellista hyötyä saavutetaan, on käsittelyyn tarvittavien laitteiden, datan hankintaan, analysointiin ja datan käsittelyyn tarvittavan kustannuksen oltava pienempi kuin big datasta saavutettava arvo. Yrityksen big datan hyödyntämisen alkuvaiheessa onkin ehkä parempi saavuttaa tuloksia käyttämällä ulkoisia palveluja. Oman datan siirtämisen suhteen yrityksen ulkopuolelle on oltava tarkat pelisäännöt.

Big data tuo mukanaan myös mahdollisuuden tietojen väärinkäyttöön. Tietovarastoihin vapaasti pääseville IT-tukihenkilöillä saattaa tulla houkutus käyttää tietovarastosta saatavaa tietoa omiin tarkoituksiinsa. Esimerkkinä maailmalta voi olla NSAn (National Security Agency) tietovarastosta kerätyt ja vuodetut tiedot Edward Snowdenin toimesta. Hän toimi NSA:n IT-tukihenkilönä ja sai näin vapaan pääsyn NSA:n big data-tietovarastoon.

Euroopan Unionin alueella henkilökisteritietojen käsittelylle on olemassa jo nyt tiukat säännöt. Säännöstöä ollaan vielä tarkentamassa. Syynä ehkä vakoilu ja tietovuodot, jotka paljastuivat vuoden 2013 aikana. Säännöstöt ja lait eivät ole ainoa tapa suojata raharvoista yrityksen tietoa. Ohjeistuksella, tarkoituksenmukaisilla tietoihin käsiksi pääsyn oikeuksilla ja tietoa käsittelevien huolellisuudella on suuri merkitys tietojen suojaukselle. Kuluttajien omilla toimilla on mahdollisuus vaikuttaa henkilökohtaisten tietojen pitämiseen yksityisinä, jos näin halutaan.

Datan avoimuudella tulevaisuudessa on paljon hyötyjä. Yrityksillä saattaa olla mahdollisuus tarjota omaa raakadataansa avoimeen käyttöön oman toimialueensa ulkopuolelle. Tiedon erilaiset analysointimenetelmät on avaintekijä, jolla voi jalostaa lisäarvoa valtavista tietomääristä. Kaikkea liikkuvaa tietoa ei myöskään ole tarvetta tallettaa, mutta poisheitetty tietokin voi olla joissakin tapauksissa arvokasta.

Yrityksissä päätökset perustuvat tietoon. Tiedon nopea saatavuus saattaa olla merkittävä kilpailuvaltti markkinatilanteita ratkottaessa. Käytettävissä olevien tietojen analysointi oikein saattaa muodostaa lähtökohtaista etua yrityksen kilpailijoihin nähden. Vastaavasti väärät tulkinnat tiedosta, jonka alkuperää ei ole varmistettu, saattavat aiheuttaa isoja-

kin taloudellisia tappioita. Koska tiedot ovat yrityksen tärkeää pääomaa, pitää ne myös suojata mahdollisia tietomurtoja vastaan. Mikäli tietokantoihin tai tallennuksiin päästään ulkoa käsiksi, saatetaan yritykselle ja sen asiakkaille aiheuttaa suuria taloudellisia tappioita.

Analytiikalla on yrityksessä vaikea analysoida tietoa, jota ei ole saatavissa. Esimerkiksi suomalaisen tapaan palveluun tyytyväisinä, tätä ei kommentoida mitenkään. Puuttuvaa informaatiota yritykselle onkin etsittävä muilla tavoin, tuottaen analyysoijalle omat haasteensa. Samoin jos irrationaaliset määrät tietoa muuttuvat ohjaaviksi tekijöiksi yrityksen analysoinnin tuloksissa, saattavat myös lopputulokset suuntautua aivan totuudenvastaiseen suuntaan. Tietojen analysoinnissa vaaditaan myös näkemystä yrityksen toimialasta ja kokonaisuudesta, jotta ei jäädä ”viilaamaan pilkkua” sopivien tulosten etsinnässä.

Opinnäytetyössäni olen myös kohdannut big datan nopeuden. Uutta materiaalia tulee saataville joka päivä useista eri lähteistä. Hadoop-järjestelmästä tuotettiin uusi kehittyneempi versio markkinoille kirjoitusvaiheen aikana. Itselleni opinnäytetyön teossa on auennut vain pieni kurkistus big datan tarjoamien mahdollisuuksien maailmaan.

LÄHTEET

Avoim data

http://www.suomi.fi/suomifi/tyohuone/yhteiset_palvelut/avoin_data/

Big Data Now: 2012 Edition. 2012. O'Reilly Media. Sebastopol, CA, Yhdysvallat: O'Reilly Media, Inc.

CIO Decisions, June 2013 vol23 (p.24-27). Tulostettu 17.9.2013.

<http://searchcio.techtarget.com/ezone/enterprise-CIO-Decisions>

ComputerWeekly.com October 2013 News. Luettu 6.11.2013.

<http://www.computerweekly.com/news/2240208220/Government-calls-for-more-data-scientists-in-the-UK>

Davis, K. & Patterson, D.2012. Ethics of Big Data. Sebastopol, CA, Yhdysvallat: O'Reilly Media, Inc.

The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, December 2012. Tulostettu 20.11.2013.

<http://idcdocserv.com/1414>

Hadoop Version 2: One Step Closer to the Big Data Goal. Tulostettu 7.1.2014.

http://www.datanami.com/datanami/2013-10-17/hadoop_version_2_one_step_closer_to_the_big_data_goal.html

Henkilötietolaki

<http://www.finlex.fi/fi/laki/ajantasa/1999/19990523>

Hotti,M.

Blogs.technet.com. 2012. Big Datan ja yrityksen oman liiketoimintatiedon yhdistäminen, tietovarastointi, analysointi ja raportointi. 6.11.2012. Tulostettu 26.11.2012. <http://blogs.technet.com/b/markohot/archive/2012/11/06/big-datan-ja-yrityksen-oman-liiketoimintatiedon-yhdist-228-minen-tietovarastointi-analysointi-ja-raportointi.aspx>

Hotti,M.

Blogs.technet.com. 2013. Big Data - hypeä vai hömppää? Microsoftin käytännönläheinen lähestymistapa auttaa ymmärtämään Big Datan mahdollisuudet ja yrityksen todelliset Big Data –skenaariot. 26.6.2013. Tulostettu 4.10.2013.

<http://blogs.technet.com/b/markohot/archive/2013/06/26/big-data-hype-228-vai-h-246-mpp-228-228-microsoftin-k-228-yt-228-nn-246-nl-228-heinen-l-228-hestymistapa-auttaa-ymm-228-rt-228-m-228-228-n-big-datan-mahdollisuudet-ja-yrityksen-todelliset-big-data-skenaariot.aspx>

Hurwitz, J. Nugent, A. Halper, F. & Kaufman, M. 2013. Big Data For Dummies, Hoboken, New Jersey: John Wiley & Sons, Inc.

Getting Started with Hadoop Planning Guide. Tulostettu 21.1.2014.

<http://www.intel.com/content/www/us/en/big-data/getting-started-with-hadoop-planning-guide.html>

21 polkua Kitkattomaan Suomeen. Helmikuu 2013. Tulostettu 10.11.2013.
http://www.tem.fi/files/35440/TEMjul_4_2013_web.pdf

Kide-raportti 2013. Julkaistu 15.5.2013. Tulostettu 31.10.2013.
<http://www.lvm.fi/julkaisu/4147800/kide-raportti-2013>

McKinsey:

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. Tulostettu 8.10.2013.
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

Mohanty, S., Jagadeesh, M., Srivatsa, H. 2013. Big Data Imperatives, New York, Springer Science + Business Media, Apress ebooks.

Olhorst, F. 2013. Big Data Analytics - Turning Big Data into Big Money. Hoboken, New Jersey: John Wiley & Sons, Inc.

Poikola, A., Kola, P. & Hintikka, K. 2010. Julkinen data - johdatus tietovarantojen avaamiseen. Liikenne- ja viestintäministeriön julkaisuja 2010. Tulostettu 7.12.2012.
<http://www.lvm.fi/julkaisu/1155483/julkinen-data-johdatus-tietovarantojen-avaamiseen>

Rekisteriseloste S-kanava. Tulostettu 21.9.2013.
<https://www.s-kanava.fi/web/s/s-kanavan-rekisteriseloste>

Rikoslaki

<http://www.finlex.fi/fi/laki/ajantasa/1889/1889039001>

Russell, M. 2011. Mining the Social Web. Sebastopol, CA: O'Reilly Media, Inc.

Salo, I. 2013. Big data – tiedon vallankumous. Jyväskylä, Docendo Finland Oy.

Sas: Roadmaps for the CIO. Tulostettu 22.1.2014
<http://www.sas.com/resources/asset/BigDataAnalytics-FutureArchitectures-Skills-RoadmapsfortheCIO.pdf>

Talouselämä 31.5.2013 Iso data nukkuu kaupassa. Tulostettu 10.11.2013.
 Talouselämä 23.5.2013 Näin S-ryhmä selittää asiakastietojen käyttöä hulluruoho- tapauksessa. Tulostettu 21.9.2013.
<http://lehtiarkisto.talentum.com/lehtiarkisto/>

TechTarget: Hadoop -2. Tulostettu 29.1.2014.
<http://searchdatamanagement.techtarget.com/definition/Hadoop-2>

Three-Legged Stool:

Big Data's Three-Legged Stool - Information Management Online Article By Jill Dyché
13.3.2013. Tulostettu 17.9.2013.

<http://www.information-management.com/news/big-data-three-legged-stool-10024077-1.html>

Tietosuoja.fi - Tietosuojavaltuutetun toimisto (1.). Luettu 23.9.2013.

<http://www.tietosuoja.fi>

Henkilötietojen käsittelyn ulkoistaminen, yhteiset tietojärjestelmät, verkottuminen ja niihin liittyvät sopimukset (2.). Tulostettu 23.9.2013.

http://www.tietosuoja.fi/uploads/fqfq98_1.pdf

Kameravalvonnan yksityisyyden suoja ja henkilötietojen käsittely (3.). Tulostettu 23.9.2013.

http://www.tietosuoja.fi/uploads/2lrt0dxzjo42lh_1.pdf

Warden, P. 2011. Big Data Glossary. Sebastopol, CA: O'Reilly Media, Inc.

Yle uutiset Kotimaa 30.10.2011 Kaupan asiakasrekisterit avattiin botulismitapausten estämiseksi. Tulostettu 21.9.2013.

http://yle.fi/uutiset/kaupan_asiakasrekisterit_avattiin_botulismitapausten_estamiseksi/5444825