OULUN AMMATTIKORKEAKOULU

Minna- Sisko Mustonen

SUURTEN TIETOMASSOJEN KÄSITTELY

Big Data

SUURTEN TIETOMASSOJEN KÄSITTELY

Big Data

Minna- Sisko Mustonen
Opinnäytetyö
Kevät 2014
Tietojenkäsittelyn koulutusohjelma
Oulun ammattikorkeakoulu

TIIVISTELMÄ

Oulun ammattikorkeakoulu
Tietojenkäsittelyn koulutusohjelma

Tekijä: Mustonen Minna- Sisko Mustonen
Opinnäytetyön nimi: Suurten tietomassojen käsittely: Big Data
Työn ohjaaja: Niva Anu
Työn valmistumislukukausi- ja vuosi: Kevät 2014 Sivumäärä: 45+1

Tämän opinnäytetyön tarkoituksena oli luoda tietopaketti tietojenkäsittelyn opiskelijoille suurten tietomassojen käsittelyyn liittyen Big Datasta. Työn toimeksiantajana toimi Oulun ammattikorkeakoulu, Liiketalouden yksikkö.

Opinnäytetyön tarkoituksena oli selvittää ilmiötä Big Data kuten mitä se on, missä ja miten sitä syntyy ja miten ja missä sitä käytetään sekä käytettävistä menetelmistä. Työssä käytiin läpi käsitteet tiedon varastointi, tiedonlouhinta, tiedon analysointi ja tiedon jalostaminen. Big Data työvälineistä esiteltiin pintapuolisesti Hadoop ja sen tärkeimmät osaprojektit. Tietoturva ja yksityisyys kohdassa käsiteltiin läpi yksityisyydensuojaa ja henkilötietoihin liittyvää perusasiaa, jotka ovat tärkeä osa Big Data- ilmiötä. Työssä käytiin myös lyhyesti läpi pilvipalvelut ja miten ne liittyvät ilmiöön. Työn lopussa tarkasteltiin muutamien case- esimerkkien kautta miten ja missä Big Dataa käytetään Suomessa ja muualla maailmassa.

Tietoperustana työssä on käytetty aiheesta olevaa kirjallisuutta, elektronisia julkaisuja ja lehtiartikkeleita. Kaikki käytetyt lähteet liittyvät Big Dataan.

Työn tuloksena syntyi pienimuotoinen tietopaketti, jota voidaan käyttää aiheeseen tutustumisessa. Aihe osoittautui laajemmaksi kuin aluksi näytti. Aiheen laajuuden vuoksi osa-alueet on jouduttu rajaamaan suppeammiksi ja niitä ei ole voitu tarkastella tässä työssä syvällisemmin. Aihetta voitaisiin tarkastella yksityiskohtaisemmin eri työvälineiden kohdalta. Tietoturva, yksityisyys ja pilvipalvelut ovat myös aihealue jota voitaisiin tutkia laajemmin.

Asiasanat:

Big Data, Hadoop, 3V malli, Yksityisyydensuoja, Henkilötietolaki, pilvipalvelut, tiedonhallinta

ABSTRACT

Oulu University of Applied Sciences
Degree Programme in Business Information Systems

Author: Mustonen Minna-Sisko
Title of Bachelor´s thesis: Processing of large information masses: Big Data
Supervisor: Niva Anu
Term and year of completion: Spring 2014 Number of pages: 45+1

This Bachelor's thesis examines processing of large information masses related to Big Data. The aim of this thesis was to create an information package about what is big data, where and how its born, what are the sources and where and how to use it. The thesis was commissioned by the Oulu University of Applied Sciences, the School of Business and Information Management.

The Purpose of this thesis was to study the phenomenon of Big Data, such as what it is, where and how it is generated as well as how and where it is used and the available methods. The thesis examines the concepts of data warehousing, data mining and data analysis and data processing. Of the Big Data tools Hadoop and its main sub-projects are presented briefly. The section of Security and privacy deals with the protection of privacy and personal data related to the basic things that are an important part of the big data phenomenon. The work also includes in a brief review of cloud services, and how they relate to the phenomenon of Big Data. The end of the thesis looks at a few examples of how and where big data is used in Finland and elsewhere in the world.

The information of this thesis is based on the existing literature, electronic sources and articles in various journals. All the sources are in some way related to Big Data.

The result of this thesis was a small-scale information package that can be used as an introduction to the topic. Subject turned out be broader than it first seemed. It was necessary to limit the scope of the subject areas and it was not possible to study these areas more thoroughly. A further research could be to examine the different tools in more detail. Security, privacy and cloud services are also topics that could be examined more extensively

Keywords:

Big Data, Hadoop, 3V model, protection of privacy, Personal Data Act, cloud service, data management

SISÄLLYS

1 JOHDANTO

Tänä päivänä tietoa kertyy valtavia määriä, valtavalla vauhdilla. Laitteiden kuten älypuhelimien ja tablettien määrä kasvaa koko ajan kiihtyvällä vauhdilla. Kehittynyt teknologia mahdollistaa kuvien, ääni- ja videotallenteiden lähettämisen sekunneissa ympäri maailmaa. Jätämme itsestämme koko ajan paljon tietoa erilaisiin järjestelmiin erilaisina klikkaus- ja lokitietoina.

Yritykset keräävät ihmisistä paljon erilaista tietoa erilaisiin rekistereihin. Tietoa kerääntyy myös erilaisista yritysten omista sisäisistä järjestelmistä. Valtavia määriä dataa syntyy myös kauppaliikkeiden ja vakuutuslaitosten ja pankkien kerätessä tietoa asiakkaista erilaisten kanta-asiakkuuksien, kuten bonuskorttien kautta, voidakseen profiloida asiakkaita ja heidän ostokäyttäytymistään tarjotakseen oikeanlaisia palveluita ja tuotteita erilaisille asiakasryhmille.

Tämän työn tarkoituksena on selvittää, mitä on Big Data, mitkä ovat Big Datan lähteet, missä ja miten sitä syntyy. Työn alussa määritellään ilmiön datan tyypit ja mitä käsite Big Data tarkoittaa. Työssä käydään läpi käsitteet tiedon varastointi, tiedonlouhinta ja tiedon analysointi ja jalostaminen. Työkalut ja välineet Big Data-massojen käsittelyyn- kappaleessa käydään läpi lyhyesti myös Hadoopin ekosysteemi ja sen tunnetuimpia osaprojekteja.

Tietoturva ja yksityisyys- kappaleessa käydään lyhyesti läpi tietoturvan määrite ja yksityisyyden suojaa ja henkilötietoihin liittyvää perusasiaa, jotka ovat tärkeä osa big data aiheessa. Pilvipalvelu- kappaleessa määritellään lyhyesti mitä pilvipalvelut ovat, ja mitkä ovat niiden etuja ja mahdollisuuksia. Työssä esitellään myös, miten ilmiö näkyy Suomessa ja muualla maailmassa muutaman case-esimerkin kautta ja käydään läpi myös muutamia esimerkkejä mahdollisuuksista ja haasteista.

2 BIG DATA- ILMIÖ

Nopeasti kasvavan ja monipuolistuvan datan laajuus muodostaa haasteen yrityksille ja yhteiskunnalle. Haasteena on ja tulee olemaan myös se, millaisia ratkaisuja organisaatiot ja yhteiskunta pystyvät ilmiöön tarjoamaan. (Salo 2013, 10.)Tänä päivänä dataa syntyy paljon. Verkkoon kytkettyjen laitteiden määrä kasvaa koko ajan ja samalla kasvaa myös virtaavan datan määrä. Teknologia mahdollistaa entistä tehokkaamman datan luomisen, sekä tallentumisen, ja teknologian kehittyessä datan määrä tulee kasvamaan entisestään. (Salo 2013, 11–12.)

Yritykset tuottavat hämmästyttäviä määriä dataa. On raportoitu, että pelkästään Facebook kerää 250 teratavua päivässä. Thompson Reuters News Analyticsin mukaan digitaalisen tiedon tuotanto on enemmän kuin kaksinkertaistunut, lähes miljoona petatavua, joka vastaa noin miljardia tera-tavua. Organisaatioiden kerätessä ja tuottaessa valtavia määriä tietoja, ne ovat tunnistaneet tietojen analysointien hyötyjä, mutta ovat myös kamppailleet, miten hallita hallussaan olevaa massiivista tietomäärää. Tämä taas on johtanut uudenlaisiin haasteisiin, kuten tehokkaaseen tallentamiseen ja käsittelyyn ja tietojen tehokkaaseen analysointiin. (Lublinsky, Smith & Yakubovich 2013, 1.) Kaikesta maailman datamäärästä noin 90 % on syntynyt viimeisen kahden vuoden aikana ja dataa kuvataankin liiketoiminnan ”uudeksi öljyksi” (Krushe-Lehtonen 2014, 20).

2.1 Strukturoitu, strukturoimaton ja semistrukturoitu data

Datasta voidaan louhia informaatiota, josta voidaan muodostaa tietoa. Kertynyt tieto taas lisää tietämystä. (Salo 2013, 26.) Dataa voidaan jakaa eri tyyppeihin, kuten strukturoituun, strukturoimattomaan, sekä näiden välimuotoon, semistrukturoituun dataan (Salo 2013, 25).

Strukturoitu data

Strukturoitu data viittaa yleensä sellaisiin tietoihin joille on määritelty tietynlainen pituus ja koko. Strukturoitu eli jäsennelty data sisältää numeroita, päivämääriä ja ryhmiä sanoista ja numeroista, joita kutsutaan stringeiksi, kuten asiakkaan nimi ja osoite. Jäsennelty tieto on yleensä tallennettu tietokantaan, josta voidaan tehdä kyselyjä käyttämällä kyselykielenä esimerkiksi SQL- kieltä (Structured Query Language). (Hurwiz, Nugent, Halper & Kaufman, kappale 2.) Strukturoitua dataa voidaan käsitellä perinteisin tiedonhallintamenetelmin ja sitä voidaan tallentaa perinteisiin relaatiotietokantoihin perustuviin järjestelmiin (Siiramaa 2012, 8).

Strukturoimaton data

Strukturoimaton data eli jäsentelemätön data on ennalta määrittelemätöntä. Kuten strukturoidussa datassakin, tiedon lähteenä voi olla koneen tai ihmisen synnyttämää tietoa. Se voi olla kuvia, videoita, tekstiä, sähköposteja, tutkimustuloksia, säätietoja, tieteellistä tietoa, asiakirjoja, paikka- ja sijaintitietoja tai sosiaalisen median synnyttämää tietoa. Esimerkkeinä satelliittikuva (Google Earth), joka sisältää säätietoa tai muuta tietoa, jota valtiot keräävät itselleen, tieteellistä tietoa, joka sisältää (seismistä kuvastoa, maanjäristys kuvia), ilmakehän tietoja, korkean energian fysiikkaa, valokuvia ja videoita, jotka (sisältävät turvallisuus-, valvonta- ja liikennevideoita), tutka- ja kaukoluotaintietoja, jotka voivat sisältää ajoneuvo-, meterologisia sekä meritieteellisiä seismisiä kuvauksia. Esimerkkejä ihmisen synnyttämistä tiedoista ovat yrityksissä syntyvät sisäiset tekstit, kuten asiakirjat, lokit, tutkimustulokset ja sähköpostit, sosiaalisen median tiedot, jotka on luotu sosiaalisen median alustoilla, kuten Facebook, YouTube, Twitter, LinkedIn ja Flickr tai Instagram, mobiilidata, joka sisältää tietoja kuten tekstiviestit ja sijaintitiedot, Web-sivustojen jakama ennalta suunnittelematon sisältö kuten Facebook ja Instagram. (Hurwiz ym. 2013, kappale 2.)

Semistrukturoitu data

Semistrukturoitu data on strukturoidun ja strukturoimattoman datan välimuoto. Se sisältää molempia data-tyyppejä, kuten datan yhteyteen liitetyt metatiedot videomateriaalin tai valokuvien yhteydessä. (Salo 2013, 25.)

2.2 Mitä on Big Data?

Mikkelän (2012, 11–12) mukaan Big Data- käsitteellä tarkoitetaan jäsentymätöntä sähköisen tiedon tulvaa, joka on viestien, automaattien, erilaisten digitaalisten tapahtumien ja digitaalisen median synnyttämää. Sillä tarkoitetaan siis suuria tietomassoja, jotka sisältävät erittäin paljon nopeaa, hajanaista ja vaihtelevampaa dataa kuin milloinkaan ennen. Sisällään datamassat pitävät monisisältöisiä ja monenmuotoisia elementtejä ja tietovirtoja, kuten kuvia, ääntä, tekstiä, numeroita, tilastoja ja lokitiedostoja.

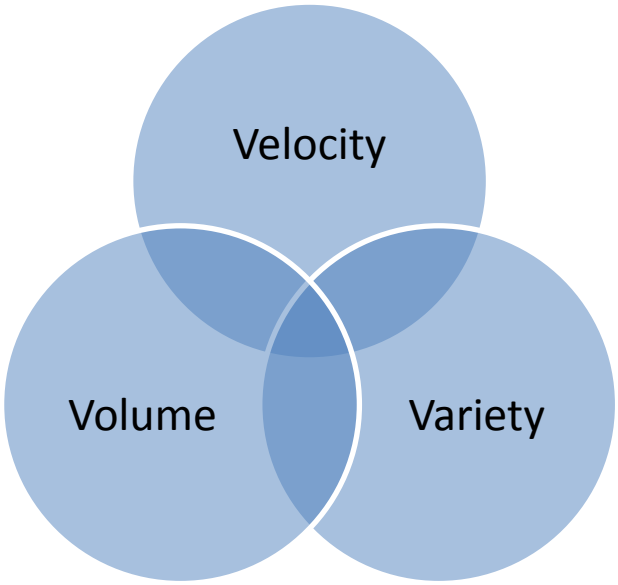
Lublinskyn ym. (2013, 1) mukaan Big Data- termiä käytetään kuvaamaan data-aineistoja, jotka ovat niin suuria ja laajoja, että tyypillisestä ja perinteisestä tietojen varastoinnista, hallinnasta, hakemisesta ja analysoinnista on tullut haaste. Big Datalle ominaista on digitaalisen tiedon koko, joka voi tulla monista eri lähteistä ja olla muodoltaan strukturoitua tai strukturoimatonta dataa. Dataa voidaan käsitellä ja analysoida, jotta voidaan löytää käsityksiä ja malleja, jotta voidaan tehdä tietoihin perustuvia päätöksiä.

Siiramaan (2012, 8) mukaan” Big Datalla tarkoitetaan missä tahansa muodossa olevaa, nopeasti muodostuvaa dataa, jota ei pystytä käsittelemään tehokkaasti perinteisillä järjestelmillä” ja ”Perinteisillä järjestelmillä tarkoitetaan relaatiotietokantoihin ja datavarastoihin (datawarehouse) perustuvia rakenteisen datan tallentamiseen ja käsittelyyn tarkoitettuja järjestelmiä” (Siiramaa 2012, 8).

Patil & Thian (2013, 3) mukaan Big Data on joukko data-aineistoja, jotka ovat niin suuria ja monimutkaisia prosesseja, että niitä on vaikeaa käsitellä perinteisillä tietokantojen hallintatyökaluilla. Haasteina ovat tiedon talteenotto, kuratointi, varastointi, etsiminen, jakaminen, siirtäminen, analysointi ja visualisointi. Big Data on yhdistelmä tiedonhallinta teknologioita, jotka ovat kehittyneet ajan kuluessa. Big Data- termiä käytetään määrittelemään suuria joukkoja dataa tai tiedostoja, jotka voivat olla tyypiltään strukturoitua, strukturoimatonta tai molempien sekoitusta, ja joka on niin nopeasti suureksi kasvavaa, että sen käsittelemisestä perinteisin tietokannoin ja tilastollisin työkaluin tulee vaikeaa.

2.3 Big Datan 3V- käsitemalli

Big Data- termi voidaan määritellä myös toisella tapaa siten, että jokaisen tietolähteen tulee täyttää vähintään kaikki 3 V-käsitemallin tunnusomaiset piirteet, jotta se voitaisiin tulkita Big Dataksi (Patil & Thia 2013, 3). Tavallisimmin Big Dataa kuvataan 3V- käsitteellä. Käsitteen 3V kirjaimet tulevat sanoista Volume eli määrä, Velocity eli nopeus ja Variety eli moninaisuus (Kuvio 1). (Salo 2013, 21–22.) Edellä kuvattujen kolmen V:n rinnalle on yhdeksi vaihtoehdoksi nostettu Big Datan neljäs V-kirjain, johtuen uudesta tuottavuus ja innovaatioryöpystä, jonka ilmiö on muodostanut (Mikkelä 2012,12). Big Datan neljäntenä V:nä voitaisiin pitää sanaa Value eli arvo. Arvo kuvastaa Big Datan tuomaa hyötyä liiketoiminnalle. (Remes 2013,9.)



KUVIO 1. Volume, Velocity, Variety (Salo 2013, 23)

Volumella eli määrällä viitataan datan suuruuteen ja sen kasvamiseen maailmassa (Salo 2013, 21–22). Volumella viitataan tietoaineistojen kokoon. Ne voivat olla kooltaan KB (kilobitti), MB (megabitti), GB (gigabitti), TB (terabitti) tai PB (petabitti) tyyppisiä sovelluksia tai kohteita, jotka tuottavat tai vastaanottavat dataa. (Prajapati 2013, kappale Introducing Big data.) Volumelle tyypillistä on datan määrä, jota syntyy jatkuvasti. Erilaiset datatyypit tulevat erikokoisina. Esimerkkinä volumesta ovat blogitekstit, jotka ovat kooltaan muutamia kilotavuja, äänipuhelut ja videotiedostot, jotka ovat kooltaan

muutamia megatavuja, kun taas anturitiedot, koneen lokitiedot ja klikkaustiedot voivat olla kooltaan gigatavuja. (Krishnan 2013, kappale 2.)

Velocityllä eli vauhdilla /nopeudella tarkoitetaan kiihtyvää nopeutta, jolla dataa syötetään ja otetaan käyttöön tietojärjestelmistä (Salo 2013, 21–22). Velocityllä voidaan määritellä kohteen nopeuden ja liikkeen suunta (Krishnan 2013, Data velocity). Velocity viittaa pienellä viiveellä, reaali-aikaiseen nopeuteen, jolla analyysit on suoritettava. Hyvin tyypillisenä esimerkkinä nopeudesta on analyysin tekeminen sosiaalisen median jatkuvalla virralla syntyvistä tiedoista tai yhdistelemällä tietoja erilaisista tietolähteistä. (Prajapati 2013, kappale Introducing Big data.)

Variety eli vaihtelevuus ja monimuotoisuus kuvastavat datan muuttumista yhä epäyhteneväisemmäksi (Salo 2013, 21–22). Vaihtelevuudella viitataan vaihtelevaan datatyyppiin, jossa voi esiintyä esimerkiksi kuvia, tekstiä, ääntä ja videoita (Prajapati 2013, kappale Introducing Big data). Dataa syntyy moninkertaisissa rakenteisissa muodoissa, vaihdellen sähköposteista sosiaalisen median twiitteihin ja sensoreiden tuottamaan dataa (Krishnan 2013, Data Variety).

2.4 Tiedon syntyminen ja -lähteet

Big dataa syntyy koko ajan valtavia määriä, erilaisista lähteistä. Tabletit, älypuhelimet, henkilökohtaiset ja teollisuuden koneet, satelliitit, erilaiset anturit ja sensorit sekä sosiaalinen media tuottavat paljon erityyppistä ja erikokoista dataa vaihtuvalla nopeudella. Dataa syntyy myös erilaisista asiakirjoista ja sopimuksista, erilaisista liiketoiminnan järjestelmistä, sähköposteista, sovellusten lokitiedoista ja klikkaustiedoista. (Krishnan 2013, kappale 2.) Erilaisia Big Datan tiedon lähteitä voivat liiketoiminta-ohjelmistojen tietokantojen ohella olla esimerkiksi sosiaalinen media, anturien ja sensorien tuottama tietovirta ja erilaiset lokitiedot (Hämäläinen 2012, 37).

Konetiedot

Kaikki päivittäin käyttämämme koneet ja laitteet, niin henkilökohtaiset kuin teollisuuden työkoneet, tuottavat paljon erilaista tietoa. Tämä data sisältää tietoa koneen käyttäjistä ja käyttäytymisestä ja se sisältää useimmiten myös yksityiskohtaisia toimintalokeja. Konepohjaiselle datalle ominaista on tasainen malli numeroita ja tekstiä, jota esiintyy nopeatempoiseen tyyliin. Esimerkkeinä konedataa synnyttävistä lähteistä ovat robotit, erilaiset anturit, radiosignaalit, satelliitit ja mobiililaitteet. (Krishnan 2013, kappale 2.)

Sovellusten lokitiedot

Konepohjaisten tietojen toinen muoto on sovellusten loki, jota eri laitteet tuottavat eri muodoissa ja eri tahtiin, kun joka sekunti internet mahdollistaa pääsyn käyttäjilleen suosittuihin käyttöjärjestelmiin miljoonien palvelimien kautta ympäri maailmaa. Erilaiset tietokonetomografilaitteet, röntgenlaitteet, lentoasemien henkilöskannerit, laivat, aseteollisuus ja kaupalliset satelliitit tuottavat dataa. Tabletit, matkapuhelimet ja autojen tietokoneet voivat tänä päivänä kaikki tuottaa lokitietoa laitteiden toiminnoista, sisältäen esimerkiksi maantieteellistä tietoa, tietotyypin, käyttöoikeustyypin ja toiminnon ajankohdan, mihin päivänaikaan tahansa. (Krishnan 2013, kappale 2.)

Klikkaustietojen lokitiedot

Tyypillinen klikkaustietojen loki tulee internetin portaaleista ja sivuilta. Klikkaustietojen data otetaan kiinni sivustoilta tilastointia varten. Data tarjoaa tietoa siitä, mitä käyttäjä tekee sivustoilla ja näin voidaan saada erittäin hyödyllistä tietoa käyttäytymisen ja käytettävyyden analysointiin, markkinointiin ja yleiseen tutkimustyöhön. (Krishnan 2013, kappale 2.)

Ulkoiset tai kolmannen osapuolen tiedot

Tänä päivänä on paljon data-aineistoja, joita organisaatiot ostavat tai saavat syötteinä ulkoisista lähteistä. Vaikkakin osa datasta on strukturoitua eli jäsenneltyä, niin suurin osa datasta on strukturoimatonta eli se useimmiten tulee eri muodoissa suurella ja raskaalla volyymilla. Esimerkkinä tästä ovat säätiedosta syntyvä data, jota erilaiset anturit tuottavat. (Krishnan 2013, kappale 2.)

Sähköpostit ja sopimukset

Yrityksissä tuotetaan päivittäin valtavia määriä erilaista tietoa. Sitä syntyy esimerkiksi työntekijöiden, avainhenkilöiden ja asiakkaiden lähettämistä sähköposteista ja erilaisista sopimuksista. (Krishnan 2013, kappale 2.)

Maantieteelliset tietojärjestelmät ja paikkatiedot

Moni älypuhelin lisää kuviin paikkatiedot automaattisesti. Glopal Positioning Systems (GPS) on suosittu laite- ja älypuhelinsovellus, joka käyttää paikkatietoja ohjatakseen kenet tahansa paikasta A paikkaan B. Lisättyjen GPS kameroiden ja videokameroiden ominaisuuksien avulla voidaan myös

asettaa sijainnit, siitä missä kuvat on otettu, yhdessä päivämäärien ja muiden tietojen kanssa. Tämä on erityisen suosittu trendi journalistien keskuudessa. Tällaisten reaaliaikaisten tietojen vuorovaikutus vaatii suuret määrät dataa edestakaisin välitettäväksi, joka sekunti miljoonille kuluttajille ympäri maailmaa. (Krishnan 2013, kappale 2.)

Sosiaalinen media

Sosiaalisen median tietojen hyödyntäminen on kiinnostavinta dataa useille yrityksille. Sosiaalisen median sivut eivät ainoastaan tarjoa kuluttajien näkökulmia asioista, vaan myös kilpailuaseman ja trendejä. Kuluttajat tuottavat joka minuutti valtavia määriä tietoa, joista saadaan erittäin tärkeitä näkemyksiä valinnoista, mielipiteistä, vaikutuksista, suhteista, merkkiuskollisuudesta ja muista asioista. Yritykset käyttävät sosiaalisen median sivuja henkilökohtaiseen tuotteiden ja palveluiden markkinoimiseen asiakkailleen (Krishnan 2013, kappale 2.) ”Sosiaalinen media muodostaa merkittävän tietovarannon ja – lähteen” (Mikkelä 2012, 12).

Anturidata

Antureiden tuottama data on esimerkki siitä nopeudesta ja vauhdista, jolla dataa tulee erilaisista antureista, kuten GPS, rakennusten lämmitys- ja jäähdytysjärjestelmät, mobiililaitteet, biometriset järjestelmät, tekniset ja tieteelliset sovellukset ja lentokoneiden anturit ja moottorit. Antureiden verkkoon tuottaman datan määrä voi vaihdella muutamista gigabiteistä terabitteihin sekunnissa. Esimerkkinä lentokoneen moottoreiden anturit tuottavat 650 teratavua dataa matkalla Lontoosta New Yorkiin. (Krishnan 2013,kappale 2.)

Matkapuhelinverkot

Yhtenä Big Datan lähteenä toimivat matkapuhelinverkot. Ne ovat suosituin tapa jakaa kuvia, musiikkia ja tietoa mobiililaitteiden välityksellä. Valtava määrä dataa, joka lähetetään mobiiliverkkoihin, antaa käsityksen verkkojen suorituskyvystä. Pelkkä tiedon määrä, jota välitetään mobiiliverkkojen kautta, antaa käsityksen esimerkiksi verkkojen suorituskyvystä, prosessoidun tiedon määrästä jokaisessa tukiasemassa, ajankohdasta, maantieteellisistä paikkatiedoista, käyttäjätiedoista, sijainnista ja datan käsittelyyn liittyvästä viiveestä. (Krishnan 2013, 2.)

2.5 Tiedonvarastointi ja tiedonlouhinta

Tietovarasto (Data Warehouse) on tietokanta, joka on suunniteltu tietojen helppoon ja nopeaan hakuun. Raakamuodossa olevaa tietoa muokataan ja jalostetaan raportointi- ja kyselykäyttöön sopivaan muotoon, jonka jälkeen se ladataan määräajoin tietovarastoon. (Hovi, Hervonen & Koistinen 2009, 14.)

Tiedon louhinta on tapa, jolla analysoidaan tietovarastossa olevia tietomassoja. Erityisillä louhintatyökaluilla pyritään löytämään tietovaraston alimman tason tiedoista tiettyjä malleja sekä tietojen välisiä riippuvuuksia. Menetelmien tavoitteena on jo olemassa olevan tiedon ominaisuuksien ja mallien mukaan ennustaa tulevaa käyttäytymistä. (Hovi, Ylinen & Koistinen 2001, 125.)

Tiedonlouhintamenetelmien (Data Mining) avulla pyritään löytämään laajoista data-aineistoista, kuten tietovarastoista, piilevää informaatiota. Menetelmien avulla datan joukosta etsitään piileviä korrelaatioita ja lainalaisuuksia, jotta voitaisiin löytää tietoa, joka helpottaa liiketoiminnan ennakoimista. Menetelmät, joita yleensä käytetään, ovat matemaattisia, tilastollisia sekä tietojenkäsittelyllisiä puoliautomaattisia analysointimenetelmiä. Niiden avulla suuria data-aineistoja työstetään algoritmisesti. Menetelmien avulla pyritään laajoista tietoaineistoista löytämään piilossa olevaa hyödyllistä tietoa. (Hovi ym. 2009, 98.)

Big Data – analytiikassa on kyse pitkäjänteisestä ja haastavasta tiedonlouhintaprosessista, sillä datan hyödyntämiskohteet, määrä, luonne ja muoto muuttuvat kokoajan ja näin ollen vaativat sovittamista ja hienosäätöä jatkuvasti. Big Datan louhinnassa pitävät muuten paikkansa samanlaiset periaatteet, käytännöt ja prosessit kuin muussakin datan louhinnassa, mutta datan luonne ja louhintaan sovellettava teknologia voivat poiketa huomattavasti organisaation aiemmista käytännöistä. (Salo 2013, 94–95.) Tiedonlouhinnalla viitataan menetelmään, jolla pyritään löytämään oleellista tietoa suurten tietoaineistojen joukosta. Louhinnan oletetaan saavan aikaan läpinäkyvyyttä, tehostavan toimintaa ja muodostavan ajantasaisia kytkentöjä, joiden avulla voidaan jatkuvasti arvioida tekemisiä ja aikaansaannoksia. (Ruckenstein 2013, 34.)

2.6 Tiedon jalostaminen ja analysointi

Yrityksiin kertyy jatkuvasti paljon monen muotoista dataa eri lähteistä. Vain pientä osaa yrityksiin kertyvästä datasta käytetään päätöksenteon tukijärjestelmissä. Dataa on verrattu uuteen öljyyn, jolla ei jalostamattomana ole käyttöä, mutta jalostettuna siitä tulee arvokasta kauppatavaraa. Hajautettu laskenta, koneoppiminen, NoSQL-tietokannat ja tietovirtojen tosiaikainen käsittely (Stream Computing) ovat Big Data- menetelmiä, joiden avulla muutamissa minuuteissa saadaan aikaiseksi analyysejä joiden tekemiseen perinteisillä menetelmillä aikaa menisi tunneista päiviin. (Hämäläinen 2012, 36–37.) Jalostettuna datasta tulee arvokas tuotannontekijä, joka tarjoaa yrityksille runsaasti liiketoimintamahdollisuuksia. Data-analytiikka on esimerkkinä yritysten keskeisimmästä kehittyvästä tavasta myynnin lisäämiseen, kulujen optimointiin, asiakaspalvelun parantamiseen sekä järjestelmien kehittämiseen, jotka ovat teknisesti aiempaa parempia. (Lång 2014, hakupäivä 6.5.2014.) Tietojen louhiminen ja seulonta ovat kehittyvän analytiikan pohjana ja oleellisinta on miten ja mihin tietoa käytetään. Big datasta tulee kehityksen myötä standarditeknologiaa. (Vänskä 2013, 10.)

Yrityksissä sekä julkishallinnon organisaatioissa tiedon jalostamisen merkitys tulee jatkuvasti kasvamaan. Tietomassa ei itsessään ole arvokasta. Jotta tietoa voitaisiin hyödyntää oikealla tavalla, pitäisi pyrkiä erottamaan tietomassasta olennainen tieto analysoitavaksi ja eteenpäin hyödynnettäväksi. (Tuominen 2013, 27.)

3 MENETELMÄT JA TYÖKALUT BIG DATAN KÄSITTELYYN

Strukturoimattoman datan hyödyntämiseen tarvitaan uudenlaisia työkaluja. Yksi tällainen työväline on Hadoop. Se on avoin ohjelmistokehys, jonka päälle voidaan rakentaa sovelluksia. Se on näistä työkaluista myös eniten kiinnostusta herättävä ratkaisu. Data voidaan muuttaa Hadoopin ja siihen liitettyjen sovellusten avulla senkaltaiseen muotoon, että sitä voidaan käsitellä perinteisillä analyysityökaluilla. (Hammarsten 2013, 14.) Hadoop soveltuu suurien datamäärien käsittelyyn. Monet suuret Big Datan käsittelyyn ja data-analytiikkaan perehtyneet ohjelmistotalot kuten Cloudera, IBM, Hortoworks ym. hyödyntävät omien Big Data -järjestelmiensä pohjana Hadoopia. (Siiramaa 2012, 23). Hadoop sisältää suuria määriä työkaluja, jotka ovat suunniteltu ja toteutettu toimimaan yhdessä. Hadoopia voidaan käyttää moniin asioihin, ja ihmiset määrittelevät usein sen perusteella, miten käyttävät sitä. (Lublinsky ym. 2013, kappale 1.) Hadoopin kyky käsitellä ja varastoida valtavia määriä dataa yhdistetään yleensä data tieteisiin. Data tieteiden eli Data sciencen tehtävänä on louhia datasta merkityksiä. Data sciencents- työ perustuu matematiikkaan, tilastolliseen analyysiin, hahmotunnistukseen, koneoppimiseen, suurteholaskentaan (tietokone), tietovarastointiin ja paljon muuhun. (Lublinsky ym. 2013, kappale1.)

3.1 Hadoop

Hadoop on avoimen lähdekoodin ohjelmistoprojekti (ohjelmistokehys) Big Datan kustannustehokkaaseen käsittelyyn. Sen tarkoituksena on luoda palvelinklusteri, jota voidaan käyttää suuren ja monimuotoisen datan tallentamiseen ja analysointiin. Se on alustariippumaton ja joustava tuoden näin ollen kustannussäästöjä. Ohjelmistokehyksen laskennallinen tehokkuus muodostuu datan käsittelyn hajauttamisesta suuresta määrästä tavallisista palvelimista muodostuvaan klusteriin, johon data tallennetaan oletusarvoisesti kolmena kopiona, joka tuo toimintavarmuutta. (Salo 2013, 80–81; Siiramaa 2012, 2). Hadoop on laajamittainen, laajaeräinen tietojenkäsittelyjärjestelmä, joka käyttää laskentaan MapReducea ja tiedon varastointiin HDFS- tiedostojärjestelmää, jotka ovat Hadoopin ydinprojekteja. (Patil ym. 2013, 3; Salo 2013, 80–81). Hadoop muodostuu neljästä ydinkomponentista,

jotka ovat Hadoop Distributed File System (HDFS), Hadoop MapRedeuce, Hadoop Common ja Hadoop YARN (Siiramaa 2012, 23).

3.1.1 HDFS

Hadoop Distribudet File System (HDFS) on toinen Hadoopin ydinprojekteista (Salo 2013, 82). HDFS on Hadoopin toteutus hajautetusta tiedostojärjestelmästä, joka hallinnoi klusterissa sijaitsevaa dataa. Se on suunniteltu pitämään hallussaan suuria määriä dataa ja hankkimaan oikeuden käyttää tätä dataa jakamiseen asiakkaille verkossa. Järjestelmällä voidaan hajauttaa suuria määriä tietoa klusterin tietokoneille. Data kirjoitetaan vain kerran, mutta luetaan monta kertaa analytiikkaan. (Lublinsky ym.2013, kappale 1.) HDFS on tiedostojärjestelmä, jota käytetään Hadoopissa. Järjestelmän data tallennetaan klusteriin, joka muodostuu useista palvelimista. Tallennettava data jaetaan lohkoihin ja lohkot hajautetaan Hadoop-klusterin palvelimille, mikä mahdollistaa datan tehokkaamman käsittelyn ja mahdollistaa myös datan rinnakkain käsittelyn. (Siiramaa 2012, 25.) HDFS on kaikkein luotettavin hajautettu tiedostojärjestelmä. Järjestelmä rikkoo tiedostot suuriin, vähintään 64 MB: n paloihin, joissa jokainen lohko kopioidaan kolme kertaa. (Patil ym. 2013, 3.) Jokainen lohko toistetaan useita kertoja, niin ettei mikään yksittäinen vika kiintolevyasemassa tai palvelimessa saa aikaan tietojen katoamista (Capriolo, Wampler & Rutherglen 2012, kappale 1). Jopa tuhansilla palvelimilla oleva levytila voidaan virtualisoida yhdeksi hakemistorakenteeksi, johon voidaan kirjoittaa dataa (Hotti 2012, hakupäivä 12.5.2014). HDFS on edullinen ja toimintavarma tallennuspaikka datalle (Salo 2013, 82).

HDFS arkkitehtuuri perustuu Googlen suunnittelemaan Google File Systeemiin (GFS). HDFS on suunniteltu levittämään dataa suurille määrille koneita ja tukemaan paljon suurempia tiedostokokoja verrattuna hajautettuihin tiedostojärjestelmiin kuten NFS. (Lublinsky ym. 2013, kappale 2.)

3.1.2 MapReduce

Hadoop MapReduce on YARN-pohjainen järjestelmä suurten tietomäärien rinnakkaiseen käsittelyyn (hadoop.apache.org, hakupäivä 27.5.2014). MapReduce on Hadoopin ydin. Se on ohjelmointimalli (laskentamalli), jonka avulla pystytään käsittelemään ja analysoimaan suuria datamääriä tavallisista palvelimista muodostuvista klustereista sekä hajottamaan niihin kertynyt datamassa pieniksi palasik-

si, jotka jaetaan klusterin muodostavaksi solmuiksi. (Schmidt & Phillips 2013,1-2; Siiramaa 2012, 29). MapReduce on lineaarisesti skaalautuva ohjelmointimalli, joka sisältää kaksi perustoimintoa Map eli karttatoiminnon ja Reduce eli vähennystoiminnon, joista kukin määrittelee avainarvoparin toisilleen. Molemmat toiminnot lähetetään samaan Reduce-toimintoon. (White 2012, kappale 2 Analyzing the data with Hadoop; Capriolo ym. 2012, kappale 1.)

Esimerkkinä MapReducen käytöstä on säätietojen louhiminen. Sääanturit keräävät tietoja joka tunti ympäri maailmaa ja kokoavat suuria määriä lokitietoja, jotka ovat semistrukturoitua ja tallennepainotteista dataa ja näin ollen hyvä ehdokas analysoitavaksi MapReducella. (White 2012, 2.)

3.2 Pig

Pig on Apachen avoimen lähdekoodin projekti. Se on tietovirtakieli, jota käytetään suorittamaan data-analyyseja. (Schmid ym. 2013, kappale 4.) Pig on yksi Hadoopin oheisprojekteista. Se on skriptikieli, joka on tarkoitettu suurten aineistojen tutkimiseen ja analysointiin. Se koostuu kahdesta osasta Pig-nimisestä suoritusympäristöstä ja Pig Latin nimisestä ohjelmointikielestä, jota käytetään ilmaisemaan tietovirtaa. (White 2012, kappale 1; Siiramaa 2012, 48). Pig toimii moottorina toteutettaessa tietovirto-ja rinnakkain Hadoopissa. Siinä hyödynnetään HDFS-tiedostojärjestelmää sekä MapReduce-käsittelyjärjestelmä. Kieli sisältää operaattorien monet perinteiset datatoiminnot, kuten liittää, lajitella ja suodattaa. Käyttäjät voivat kehittää omia toimintojaan lukemisessa sekä käsitellä ja kirjoittaa tietoa. Gates 2011, 1)

3.3 Hadoopin muut oheisprojektit

Hive

Hive on tietovarastoinfrastruktuuri, joka on rakennettu Hadoopin päälle. Sitä käytetään tietojen yhteenvetoihin, analysointiin ja ad-hoc kyselyihin. Hive luotiin alun perin Facebookiin, mutta siirrettiin myöhemmin takaisin osaprojektina Hadoopin ekosysteemiin. Sitä ei ole suunniteltu käsittelemään suoria transaktioita eikä synnyttämään reaaliaikaisia tuloksia. (Jain 2013, kappale 1.) Hive mahdollis-

taa SQL-tyyppisten kyselyjen käyttämisen datan hakemiseen ja analysoimiseen HDFS-tiedostojärjestelmästä. Se tulkitsee HiveQL-kielellä kyselyt MapReduce-ajoiksi, eikä käyttäjän tarvitse kirjoittaa Mapper-tai Reducer-funktioita itse. (Salo 2013, 85.) Hive muodostuu kahdesta osasta Hive-suoritusympäristöstä sekä HiveQL-kyselykielestä. HiveQL on perinteisen SQL-kielen kaltainen kyselykieli, jota käytetään Hivessä. Suoritusympäristössä Hive-kyselyt muutetaan MapReducen Map- ja Reduce-tehtäviksi. Tietojen hakeminen on kuitenkin hitaampaa kuin se olisi haettaessa perinteisitä tietokannoista ja prosessi voi kestää jopa useita minuutteja. (Siiramaa 2012, 8, 49.)

HBase

HBase on Hadoop-tietokanta Big Datan varastointiin. HBase on tietokantakomponentti, joka on rakennettu Hadoopin päälle. Se on hajautettu NoSQL-järjestelmä, joka ei sisällä kehittyneiden sarakkeiden tietotyyppejä tai toissijaisia indeksejä. Se soveltuu satojen miljoonien tai miljardien rivien tallentamiseen.(hbase.apache.org, hakupäivä 2.6.2014.) HBase on avoimen lähdekoodin toteutus Googlen Big Table arkkitehtuurista. Se on samankaltainen kuin perinteinen relaatiotietokantojen hallintajärjestelmä RDBMSa. HBasessa tiedot järjestetään taulukoissa. (Lublinsky ym. 2013, kappale 2.) HBase on jaettu skaalautuva tietovarasto, joka tukee rivitasoisia tapahtumia eli transaktioita. Yksi sen tärkeimmistä ominaisuuksista on ylläpitää sarakepainotteista (column-oriented) varastointia, jossa sarakkeet (columns) voidaan järjestää sarakeperheisiin (column families). Sarakeperheet ovat tallennettuina fyysisesti yhdessä hajautetussa klusterissa, joka tekee lukemisesta ja kirjoittamisesta nopeampaa, koska tyypillinen kyselytapahtuma sisältää vain pienen osajoukon sarakkeita. HBasea käytetään HDFS:n tietojen tallentamiseen ja sen tarkoituksena on käyttää datan in-memory muistia ja paikallisia tiedostoja lokitietojen liittämistä varten. (Capriolo ym. 2012, kappale 1.)

Zookeeper

Zookeeper on Apache Software Foundationin hallinnoima avoimen lähdekoodin projekti. Zookeeper on sovelluskirjasto, jossa on kaksi pääasiallista toteutusta API- Java ja C ja huoltopalvelukomponentti. Se mahdollistaa koordinointitehtävät hajautettuihin järjestelmiin. Koordinointitehtävä on tehtävä, johon sisältyy useita prosesseja. (Junqueira & Reed 2013, kappale 1.) Zookeeper on Hadoop- kluste-

rin palvelin. Se on työkalu, joka on tarkoitettu Hadoop- klusterin keskitettyyn konfiguraation hallintaan sekä solmujen tilan synkronointiin. Sen avulla voidaan esimerkiksi hallita nimipalveluita, ryhmäpalveluita ja synkronointipalveluita ja ylläpidetään klusterin nimiavaruushierarkiaa ja konfiguraatioasetuksia. (Siiramaa 2012, 50.)

Sqoop

Sqoop on Apache Hadoopin huipputason projekti ja se on suunniteltu siirtämään dataa Hadoopin ja RDBMS:n välillä (Jain 2013, 1). Sqoop on komentorivityökalu, jolla voidaan yhdistää relaatiotietokantoja ja Hadoop (Zburivsky 2013,kappale 3). Sqoop tarjoaa tavan tuoda ja viedä tietoa relaatiotietokannan tauluista kuten SQL Server ja HDFS. Työkalulla on kaksi tärkeää piirrettä: Import ja Export. Sqoop import tarkoittaa perinteisen RDBMS datan lataamista Hadoopiin, HBaseen ja Hiveen. (Jain 2013, 1.)Tietojen tuonti relaatiotietokannoista Hadoopiin ja toisin päin on hyvin yleinen tehtävä työkalulle. Se on melkein aina käytössä Hadoop- klusterin ulkopuolella ja sen tulevaisuuden versiot tulevat tarjoamaan asiakas-palvelin tyyppisiä yhteyksiä. (Zburivsky 2013, kappale 3.)

Impala

Impalan tehtävänä on hakea reaaliaikaisia kyselyjä Hadoopille. Impala koostuu useista komponenteista. Impala on kirjoitettu pääasiassa C++-kielellä. Impala on suunniteltu yhteensopivaksi Hiven kanssa. Se käyttää samaa HiveQL- kieltä ja hyödyntää myös jopa olemassa olevia Hive Metastorea saadakseen tietoa taulukoista ja sarakkeista. (Zburivsky 2013, 3).

R

R on avoimen lähdekoodin ohjelmistopaketti tilastollisten tietojen analyysien tekemiseen. R on ohjelmointikieli, jota käyttävät datatieteilijät sekä muut käyttäjät, jotka tekevät tilastollisia analyyseja datasta. R tarjoaa laajan ja monipuolisen koneoppimisen, johon kuuluu lineaarisesta ja ei- lineaarisesta mallintamisesta, klassisia tilastollisia testejä, aikasarja-analyyseja, luokittelua ja klusterointia ja laajennettavissa olevia graafisia tekniikoita. R:ssä on sisäänrakennettuina niin valmiit kuin lisätoiminnot tilastollisia, koneoppimis- ja visualisointitehtäviä varten, esimerkkeinä tiedonlouhinta, datan puhdistaminen ja datanlatausjärjestelmä, datan muutos, tilastollinen analyysi, ennustava mallinnus ja datan visualisointi. (Prjapati 2013, introducing R.)

4 TIETOTURVA JA YKSITYISYYS

Big Data-analyysien tullessa yrityksiin yhä tärkeämmäksi asiaksi tulee tietoturva ja datan suojaaminen (Hurwiz ym. 2013, kappale 1). Kaikista toimista sähköisessä ympäristössä tallentuu tietoa, joka saattaa mahdollistaa Big Data- sovellusten avulla henkilön yksityisyyden suojaan kajoamisen. Tiedot, jotka tallentuvat automaattisesti eri järjestelmiin mahdollistavat sen, että tiedonhaltijoilla voi olla paljon henkilökohtaisia tietoja henkilöistä. Data, jota sovelluksissa hyödynnetään, sisältää yleensä sellaista tietoa, joka täyttää henkilötiedon määritelmän, joka on asetettu henkilötietoja koskevassa lainsäädännössä. Tietosuojalainsäädännön tiukan ja yksityiskohtaisen säätelyn piiriin kuuluvat kaikki tietokannat ja tietojärjestelmät, jotka sisältävät tietoa, jonka perusteella henkilö voidaan tunnistaa. (Lång 2014, hakupäivä 6.5.2014.)

Sosiaalisen median palvelut, kuten Facebook ja Twitter, leviävät voimakkaasti. Ne ovat esimerkki alueesta, jolla kokonaan uudenlaisia henkilötietoihin liittyviä ongelmia syntyy. Suurten tietokantojen avaaminen open data-hengessä luo mahdollisuuksia uudenlaisille innovaatioille sekä palveluille, mutta saattaa hämärtää myös vastuukysymyksiä, jotka liittyvät henkilötietojen suojaan. Esimerkiksi urheilusuorituksia tai terveydentilaa koskevaa numeerista dataa voidaan käyttää parantamaan tietoisuutta omasta tilasta. Näitä tietoja matkapuhelinsovellusten avulla erilaiset sensorit voivat tarjota käyttäjilleen. Tiedot ovat henkilötietoja ja niiden käsittelyyn voi liittyä ongelmia, mikäli ne eivät pysy yksin asianomaisen kontrollissa. (Pitkänen, Tiilikka & Warma 2013, 2.)

Esimerkiksi terveydenhuoltoalalla potilastietojen ja henkilöiden yksityisyyden suojaamisessa tulee ottaa huomioon, ketkä saavat nähdä potilastietoja ja missä olosuhteissa. Tämän tyyppiset turvallisuusvaatimukset on otettava huomioon jo sovelluksia tehdessä. (Hurwiz ym.2013, kappale 1.)

4.1 Tietoturvan määrite

Tietoturvalla ja tietoturvallisuudella tarkoitetaan tietojen, palveluiden, järjestelmien ja tietoliikenteen suojaamista ulkopuolisilta. Tietoturvan tärkeimpiä tavoitteita ovat luottamuksellisuus, eheys, kiistämättömyys, saatavuus, pääsynvalvonta ja tarkastettavuus. Tietojen sekä dokumenttien turvaluokitus määrittelee oikeuden niiden käyttöön, säilytykseen ja hävittämiseen. (Suomen Internetopas, hakupäivä 25.3.2014.)

4.2 Yksityisyyden suoja

Tietosuojalla tarkoitetaan henkilöön sekä hänen toimintaansa liittyvien tietojen suojaamista luvatonta käyttöä ja keräämistä vastaan. Sen kohteena on aina luonnollinen henkilö eli ihminen. (Järvinen 2010, 15.)Tietosuojalla pyritään takaamaan ihmisille oikeus yksityisyyteen. Sillä pyritään myös estämään tietojen epäasiallinen ja tarpeeton käyttö. (Järvinen 2012, 12.) Tietosuojan tarkoituksena on turvata tiedon kohteen yksityiselämää, oikeuksia ja etuja. Se on perustuslaillinen oikeus, jonka tehtävänä on turvata ihmiselle oikeus elää elämäänsä ilman kenenkään aiheetonta puuttumista siihen. (Andersson, Koivisto & Ylipartanen 2013, 14.)

Tietoturvan tarkoituksena on pyrkiä säilyttämään kerätyt tiedot eheinä, luottamuksellisina ja saatavilla ja sen kohteena on tieto itse (Järvinen 2010, 15). Tietoturvassa suojataan tietoja ja tietojärjestelmiä, sekä pyritään varmistamaan käytön turvallisuus ja järjestelmien toiminta kaikissa olosuhteissa (Järvinen 2012, 12). Tietosuojan näkökulmasta tietoturvalla tarkoitetaan sellaisia toimenpiteitä, joilla pyritään turvaamaan ja suojaamaan henkilön yksityisyyttä, etuja ja oikeuksia. Tällaisia toimenpiteitä ovat tiedon laadun, eheyden ja luottamuksellisuuden säilyttäminen ja suojaaminen sekä teknisin että organisatorisin menettelytavoin.(Andersson ym. 2013, 14.)

4.3 Henkilötietolaki

Lain tarkoitus ja soveltaminen

Henkilötietolain tarkoituksena on toteuttaa yksityiselämän suojaa ja muita yksityisyyden suojaa turvaavia perusoikeuksia henkilötietoja käsiteltäessä. Lain tarkoituksena on myös hyvän tietojenkäsittelytavan kehittämisen ja noudattamisen edistäminen. (HeTil 1 §.)

Henkilötietojen määrittely

Henkilötiedolla tarkoitetaan kaikenlaista luonnollista henkilöä tai hänen ominaisuuksiaan tai elinolosuhteitaan kuvaavia merkintöjä, jotka voidaan tunnistaa häntä, hänen perhettään tai hänen kanssaan yhteisessä taloudessa eläviä koskeviksi (HeTil 3 § 1k).

Henkilötietojen käsittely

Henkilötietojen käsittely määritellään henkilötietolaissa. Henkilötietojen käsitellyllä tarkoitetaan kaikkia henkilötietoihin liittyviä toimenpiteitä, kuten henkilötietojen kerääminen, tallettaminen, käyttäminen, järjestäminen, luovuttaminen, siirtäminen, säilyttäminen, muuttaminen, poistaminen, yhdistäminen, suojaaminen ja tuhoaminen sekä muita henkilötietoihin kohdistuvia toimenpiteet. (HeTil 3 § 2 k.)

Yleiset periaatteet käsittelylle

Henkilötietojen käsittelyn yleisiin periaatteisiin kuuluu, että henkilöstä saadaan rekisteröidä vain käyttötarkoituksen kannalta tarpeellisia tietoja. Tarpeellisina käsittelyn kannalta tietoja voidaan pitää kun ne ovat olennaisia ja asian mukaisia ja niitä käytetään siihen tarkoitukseen kuin ne alun perin on kerätty, lukuun ottamatta historiallista tutkimusta, tieteellistä tarkoitusta ja tilastointia. (Karttunen, Laasanen, Sippel, Uitto & Valtonen 2012, 93.)

Arkaluonteiset tiedot

Henkilötietolaissa arkaluontoisiksi tiedoiksi määritellään henkilötiedot, jotka kuvaavat tai tiedot jotka on tarkoitettu kuvaamaan henkilön rotua tai etnistä alkuperää, yhteiskunnallista, poliittista tai uskonnollista vakaumusta tai ammattiliittoon kuulumista, rikollista tekoa, rangaistusta tai muuta rikoksen seuraamusta, henkilön terveydentilaa, sairautta, vammaisuutta, sekä henkilöön kohdistettuja hoito-

toimenpiteitä tai niihin verrattavia toimia. Arkaluontoisiksi tiedoiksi luetaan myös tiedot, jotka kuvaavat henkilön seksuaalista suuntautumista ja käyttäytymistä sekä sosiaalihuollon tarvetta, saatuja palveluja, tukitoimia ja etuuksia. (Karttunen ym. 2012, 94.)

Henkilötunnukset

Henkilötunnus eli sosiaaliturvatunnus on yksilöimiskeino, jota käytetään kansalaisten tarkkaan yksilöimiseen. Suomessa henkilötunnuksen saa syntymätodistuksen perusteella Suomessa tai ulkomailla syntynyt Suomen kansalainen. (Väestörekisterikeskus, hakupäivä 25.3.2014.) Henkilötunnusten käsittelystä on säädetty laissa. Niitä saa kerätä, tallettaa ja käsitellä ainoastaan laissa säädetyllä tavalla. Henkilötunnusten käsittelyyn vaaditaan aina rekisteröidyn eli henkilön suostumus. Rekisteröidyn suostumuksesta huolimatta, rekisterinpitäjän on huolehdittava, ettei tunnusta merkitä henkilörekisterin perusteella laadittuihin dokumentteihin. (Karttunen ym. 2012, 94.)

Henkilötietojen siirto Euroopan unionin ulkopuolelle

”Henkilötietoja voidaan siirtää Euroopan unionin jäsenvaltioiden alueen tai Euroopan talousalueen ulkopuolelle ainoastaan, jos kyseisessä maassa taataan tietosuojan riittävä taso” (HeTil 5 § 22 k).

4.4 Henkilörekisterit

Henkilörekisteri on henkilötietoja sisältävä joukko tietoja, jotka kuuluvat yhteen käyttötarkoituksensa vuoksi ja joita käsitellään kokonaan tai osissa. Tyypillisiä henkilörekisterejä ovat esimerkiksi yritysten asiakasrekisterit. (Karttunen ym. 2012, 95.)

> Henkilörekisterillä tarkoitetaan henkilötietolain määritelmän mukaan käyttötarkoituksensa vuoksi yhteenkuuluvista merkinnöistä muodostuvaa henkilötietoja sisältävää tietojoukkoa, jota käsitellään osin tai kokonaan automaattisen tietojenkäsittelyn avulla taikka, joka on järjestetty kortistoksi, luetteloksi tai muulla näihin verrattavalla tavalla siten, että tiettyä henkilöä koskevat tiedot voidaan löytää helposti ja kohtuuttomitta kustannuksitta. (Tietosuojavaltuutetun toimisto, hakupäivä 18.3.2014.)

Rekisterinpitäjän velvollisuudet

Rekisterinpitäjäksi määritellään se, jonka käyttöä varten henkilörekisteri perustetaan tai jonka tehtäväksi se ylläpito on laissa säädetty. Se voi olla yksi tai useampi henkilö, yritys, laitos, säätiö tai muu yhteisö.(Karttunen ym. 2012, 95.)

Huolellisuusvelvoite

Huolellisuusvelvoitteen mukaan henkilötietoja on käsiteltävä laillisesti ja käsittelyssä on noudatettava huolellisuutta ja hyvää tietojenkäsittelytapaa. Rekisterinpitäjää ja sitä käsitteleviä henkilöitä koskee vaitiolovelvollisuus sekä virheettömyysvaatimus, joka tarkoittaa, että virheellisiä, epätäydellisiä ja vanhentuneita tietoja ei saa käsitellä. (Karttunen ym. 2012, 96.)

Suunnitteluvelvoite

Suunnitteluvelvoitteen mukaan ennen kuin henkilötietoja kerätään, on määriteltävä tietojen käsittelytarkoitus, hankintalähteet ja luovutuskohteet. Rekistereiden tietoja saa käyttää ainoastaan alkuperäiseen tarkoitukseen ja toiminnalle tarpeettomat henkilörekisterit on hävitettävä. (Karttunen ym. 2012, 96–97.)

Rekisteriseloste

Suunnitteluvelvoite edellyttää laatimaan rekisteriselosteen ylläpidettävästä rekisteristä. Rekisteriselosteen tullee olla avoin ja jokaisen saatavilla ja kaikilla tulee olla oikeus vaatia sitä nähtäväkseen. Rekisteriselosteessa tule olla rekisterinpitäjän nimi ja yhteystiedot, kuvaus henkilötietojen käsittelyn tarkoituksesta, kuvaus rekisteröityjen ryhmästä ja niihin liittyvistä tiedoista ja tietoryhmistä. Siinä tulee olla myös kuvaus siitä mihin tietoja luovutetaan ja siirretäänkö niitä, sekä rekisterin suojauksen periaatteiden kuvaus. (Karttunen ym. 2012, 97.)

5 PILVIPALVELUT

Palvelut, jotka liittyvät suurten tietomassojen käsittelyyn ja hallintaan ovat liiketoiminta-alue, joka kasvaa huomattavalla nopeudella. Pilvipalvelut ovat esimerkki palveluratkaisuista, jotka tarjoavat työkaluja esimerkiksi tera-, peta-, ja eksabittien hallintaan ja käsittelyyn. (Mikkelä 2012,12.) Tulevaisuuden palveluista suurin osa tullaan toteuttamaan jonkinlaisesta pilvestä, riippumatta siitä, sijaitsevatko tiedot omissa konesaleissa vai muiden konesaleissa ostettuina palveluina (Kiiski 2013, 10). Big Data pilvipalveluna voi tarkoittaa käytännössä joko tallennusratkaisua pilvipalveluina tai analytiikkaa palveluna. Palveluntarjoajien aiemmin myymät pilvipalveluratkaisut ovat nyt Big Dataa pilvipalveluina. (Salo 2013, 103.)

Pilvipalvelut ovat Pilvipohjaisia IT-palveluita (Cloud Computing). Niillä tarkoitetaan ohjelmistoja eli tietotekniikkaratkaisuja, palveluja, laskentatehoa tai tallennuskapasiteettia, joita käytetään Internetissä. Näiden avulla yrityksissä voidaan täydentää tai korvata omia järjestelmiä. Tietoturva-huolia aiheuttaa, kun palveluja ja niissä olevia tietoja siirretään organisaation oman verkon ja hallinnan ulkopuolelle. Pilvipalvelut ovat yleensä hajautettuja, eivätkä käyttäjät tiedä minne heidän tietojaan tallennetaan fyysisesti. Tämä saattaa aiheuttaa sen että tietoja tallennetaan EU:n ulko-puolelle ja näin ollen saatetaan rikkoa esimerkiksi henkilötietojen osalta EU:n henkilötietodirektiiviä. (Andersson ym. 2013, 93–94.)

Pilvipalvelut tarjoavat joustavuutta. Loppumaton tallennustila ja laskentakapasiteetti joustavat tarpeen mukaan. Pilvipalvelut tarjoavat vaihtoehdon avointen ja maksullisten datavarantojen keskitettyyn jakeluun niin, että ne ovat yhdistettävissä organisaatioiden palveluissa säilyttämään dataan. Suurten julkisten datamassojen levittäminen laajasti vaatii keskitettyä ratkaisua, johon pilvipalvelut tarjoavat ratkaisun. Ne tarjoavat hyvän alustan julkisten datavarantojen jakamiselle kuin myös yksityisten datavarantojen myymiselle yritysten ja organisaatioiden käyttöön. Pilvipalveluiden haasteena Suomessa ovat tietoturva ja tietosuoja sekä erilaiset juridiset kysymykset, sillä tunnetuimmat palvelun tarjoajat kuten Microsoft ja Amazon eivät tarjoa datan tallennuspaikan rajaamiseen yksittäisiä valtioita, vaan takaavat datan sijaitsevan palvelinkeskuksessa, joka on EU-alueella. (Alanko & Salo 2013, hakupäivä 28.2.2014.)

5.1 Teknologiat, palveluntarjoajat ja tuotteet

Big Datassa datan määrä on niin suurta ja kirjavaa, ettei sitä voida käsitellä perinteisin menetelmin vaan sen käsittelyyn tarvitaan avuksi teknologiaa. Hadoop ja MapReduce ovat esimerkkejä valtavien datamassojen ja monimuotoisen ja vaihtelevan datan käsittelyyn soveltuvista teknologioista. Ne eivät ole vielä valmiita liiketoimintaratkaisuja, vaan ne vaativat erikoisosaamista, kokeilemista ja sovelluskehitystyötä. Big Data- teknologian pääasiallisina vaatimuksina liiketoimintatarpeiden kannalta on, että ne mahdollistavat uudenlaisen datanlouhinnan, jalostamisen sekä datasta oppimisen. Jotta voitaisiin tuottaa jatkuvasti todennettavaa arvoa, tarvitaan ymmärryksen synnyttämää kykyä prosessin automatisoimiseen, jotta voitaisiin yhdistää jalostettua Big Dataa ja analysoimaan sitä liiketoimintajärjestelmistä saatavan datan kanssa. (Salo 2013, 95.)

Elastic MapReduce on Amazon Web Services tarjoama pilvipalvelu, massiivisten tietomäärien tehokkaaseen analyysiin. Se käynnistää valmiiksi konfiguroidun Hadoop- sovelluksen. (Salo 2013, 104–105.) Amazon Elastic MapReduce on Amazon Web Service palvelu, joka antaa luvan aloittaa ja käyttää skaalautuvia Hadoop- klustereita Amazonin infrastruktuurin sisällä. MapReducea voidaan käyttää Hadoopin tavoin suurten tietomäärien analysoimiseen. (Schmidt ym. 2013, kappale 1.)

Microsoft Windows Azure HDInsight tarjoaa Big Data - tarpeisiin kattavan pilvipalvelukokonaisuus ratkaisun, jossa Hadoopin lisäksi on tarjolla valmiiksi asennetuttuna sen sisarprojekteja. Palvelu on maksullinen ja kaikkien saatavilla. (Salo 2013, 105.) Microsoft Windows Azure Insight on Hadoop-pohjainen pilvipalveluratkaisu, joka tarjoaa tietoalustan erikokoisen strukturoidun ja strukturoimattoman datan hallintaan (azure.microsoft.com, hakupäivä 12.6.2014).

Google BigQuery on Googlen tarjoama Big Data- palvelu. Siinä analysoitava data tallennetaan Google Storage nimiseen pilvipalveluun, jossa datan määrää on rajattu muutamaan teratavuun. Big Query tarjoaa kolmea eri käyttötapaa: komentorivityökalun, selainkäyttöisen käyttöliittymän ja sovellusrajapinnan eli API:n. (Salo 2013,106.) BigQuery on Big Data- tekniikka, jolla voidaan analysoida Googlen pilveen tallennettuja tietoja. Investointia omiin laitteisiin ei tarvita. (Hämäläinen 2012, 39.)

5.2 Mahdollisuudet yrityksille

Pilvipalveluina tarjottavilla ratkaisuilla etuina on, etteivät ne vaadi pitkäkestoista sopimuksellista sitoutumista. Kapasiteettitarvetta ei tarvitse tietää etukäteen, eivätkä ne vaadi investointeja ohjelmistoihin ja laitteisiin. Tallennustilan hinta pilvessä on edullista. Esimerkiksi Amazon Web Servies Glacier-palvelun tarjoama tallennustila pilvessä: teratavun tallentaminen Irlannissa sijaitsevassa palvelinkeskuksessa maksaa alle sata euroa ja petatavun tallentaminen ei ylitä sadantuhannen euron kuukausihintaa. (Salo 2013, 103.)

6 SUOMI JA MUU MAAILMA

Suomessa Big Dataa hyödynnetään tällä hetkellä vähemmän kuin esimerkiksi muissa Pohjoismaissa. Suomessa todennäköisiä Big Datan käyttäjiä tulevat olemaan sellaiset yritykset, joille on kertynyt paljon asiakastietoa, kuten kauppaliikkeet. Myös pankit, vakuutuslaitokset ja teleoperaattorit tulevat ottamaan Big Data-työkalut käyttöönsä. (Storås, 2012, 8.) ”Datan hyödyntämisen aste yrityksissä vaihtelee valtavasti”. Suomessa kaupat ja pankkisektori ovat tehneet pitkään hyvätasoista analytiikkaa. Ajoissa kehitykseen mukaan ovat heränneet myös verkko- ja mobiiliyhtiöt. Teollisuus Suomessa tulee perässä. (Vuokola 2013, 9) Big Datan tunnetuimmat sovellukset kohdistuvat tietoon, joka on ihmisten luomaa. Markkinoinnin ja myynnin analysoitaviksi hyviä kohteita ovat esimerkiksi sosiaalinen media, web-sivustojen klikkausvirrat, teleoperaattorien puhelutiedot ja luottokorttiyhtiöiden maksudata. (Hämäläinen 2012, 41.)

Tieto ja viestintätekniikan tutkimus TiViT Oy on käynnistänyt huhtikuussa 2012 Data to Intelligence-ohjelman, joka tuo yhteen yrityksiä, kymmenen tutkimuslaitosta eri toimialoilta, sekä Big Dataa jalostavien menetelmien asiantuntijoita. Niillä on liiketoiminta- sekä palvelutarpeita liittyen Big Dataan. (Hämäläinen 2012, 38.)

Muualla maailmassa Facebook, Amazon ja Google ovat esimerkkejä verkkojäteistä, jotka ovat edelläkävijöitä Big Dataan liittyen. Nämä suuret yritykset ovat myös mukana kehittämässä Hadoopia. (Storås 2012, 8.)

6.1 Esimerkkejä Big Datan käytöstä suomalaisissa yrityksissä

Case 1. Media ja toimitustyö

”Big Dataa hyödyntämällä voidaan tulevaisuudessa tehdä toimitustyötä tehokkaammin ja löytää uusia asioita. Voidaan tehdä juttuja jotka kiinnostavat lukijoita enemmän.” (Järn 2013, 9-10.) Uutisnenä on järjestelmä, joka on osa viestintäalan Next Media-hanketta. Se on kehitetty seuraamaan erilaisia sosiaalisen median palveluita kuten keskustelufoorumeita: Twitteriä sekä Facebookia ja keräämään

viestejä ja analysoimaan niitä ajan suhteen tapahtuvien muutosten varalta, jolloin voidaan huomata poikkeamia tietyissä ryhmissä ja niiden ympärillä. (Järn 2013, 9-10.)

Case 2. Kuljetusala

Helsingin Bussiliikenne on syksyllä 2011 ottanut käyttöönsä järjestelmän, joka kerää tietovarastoon valtavasti tietoa ajotavoista, jotta polttoainekustannuksia saataisiin pienemmiksi. Sekunnin välein autoihin sijoitetut dataloggerit keräävät tietoa järjestelmään siitä, mitä autoissa tapahtuu. Vuorokaudessa dataa syntyy noin 15 miljoonaa riviä, jotka siirretään päivittäin tietovarastoon ja yhdistetään yrityksen toiminnanohjausjärjestelmän tietoihin. Tietovarastoon kertynyttä voidaan hyödyntää analysoimalla nykyisiä tietoja, sekä liittämällä uusia tietoja siihen. Tietovarastossa olevista tiedoista tehdään erilaisia raportteja. Voidaan tehdä raportointi esim. polttoaineen kulutuksesta tuntitasolla, josta nähdään miten polttoaineen kulutus vaihtelee esim. ruuhka- ja yöaikana. (Siltala 2014, 12–13.)

Case 3. Peliala

Pelitalo Supercell on suomalainen pelialan yritys. Yrityksessä hyödynnetään ammattimaisesti Big Dataa. Yhtiössä työskentelee yhteensä 85 työntekijää, joista kolme on data scientisteja eli datatieteilijöitä. Useilla heistä on matemaatikon, fyysikon tai tilastotieteilijän koulutus sekä tutkijan tausta. Pelkkä akateeminen koulutus ei kuitenkaan riitä, vaan data scientistin täytyy pystyä myös kommunikoimaan. Työkaluina yrityksen datatieteilijät käyttävät Hadoopia sekä kaikkea siihen liittyvää tekniikkaa. Lopputyökaluna he käyttävät R-kieltä. ”Tasapainotamme peliä ja kehitämme sitä analyysin pohjalta. Analysoimme miten sitä pelataan, mitä miten pelaaja käyttää eri ominaisuuksia ja teemme siitä päätelmiä” kertoo Supercellin data scientist Ville Suur-Uski. (Vuokola 2013, 9.)

Case 3. Teleoperaattori

Esimerkkinä teleoperaattorien Big Datan käytöstä ovat asiakaspalveluihin tulleet puhelut, jotka voidaan analysoida. Tekniikan avulla voidaan tunnistaa eri äänenpainot, jonka perusteella pystytään tunnistamaan puhujan mielentila esimerkiksi onko hän innostunut vai tylsistynyt. Asiakaspalvelijan puheesta voitaisiin myös analysoida otoksia esimerkiksi siitä, ovatko muistaneet myydä uusia palveluja. (Storås 2012,8.) Operaattorit käyttävät Big Dataa apunaan kiinteähintaisten mobililiittymien hinnoittelussa saadakseen oikeanlaisen katteen myymilleen liittymille (Remes 2013, 9).

Case 4. Teollisuus

Softability ja ABB ovat yhdessä toteuttaneet ennakoivan huollon järjestelmän sähkömoottoreita ohjaaviin taajuusmuuttajiin. Järjestelmän tehtävänä on suurten tietomassojen tallennuksen ja analysoinnin lisäksi viedä laitteista koottu ja jalostettu data aluksi huoltotoiminnan suunnitteluun, jonka jälkeen sitä voidaan käyttää esimerkiksi tuotekehityksen tai toiminnanohjauksen tukemiseen. Laitteista kertyvää dataa pystytään prosessoimaan aiempaa monipuolisemmin ja tehokkaammin ja siihen voidaan liittää yrityksen ulkopuolista dataa. Järjestelmän pääteknologia-alustaksi valittiin Hadoop, joka toteutettiin Microsoftin Azure- pilvipalveluun. Valinnalla turvattiin alhaiset alkuinvestoinnit ja taattiin tarpeiden mukainen turvallinen ja skaalautuva ympäristö. (Manninen 2014, Hakupäivä 7.5.2014.)

Case 5. Tutkimustyö

Tieteen tietotekniikan keskuksen CSC: n tutkijat ovat soveltaneet Hadoop MapRecucea bioinformatiikassa. Polttavin Big Datan sovellusalue löytyy geenitutkimuksen uuden ajan sekvensointimenetelmistä. Tutkittavat tietomassat ovat kooltaan jopa kymmeniä teratavuja. (Hämäläinen 2012, 41.)

6.2 Esimerkkejä Big Datan hyödyntämisestä yrityksissä muualla maailmassa

Case 1. Autoteollisuus

Autojätti Ford käyttää analytiikkaa kaikessa mitä yhtiössä tehdään. Noin 200 Big Data- ja analytiikka osaajaa yhtiön eri toiminnoista viipaloivat tietoa kuluttajien mielialoista, bulkkitavaroiden hintaennusteista, komponenttitoimittajien valinnasta ja henkilö- ja kuorma-autovalikoiman optimaalisista voimalaiteratkaisuista. Esimerkkinä ”Kaikkiin Fordin myymiin hybridiautoihin on asennettu mobiiliyhteys, jonka asiakas voi ottaa käyttöön halutessaan. Se lähettää ajoneuvon toiminnasta dataa suoraan valmistajalle”. ”Tulevaisuuden auto voi olla sensorialusta, joka kommunikoi tien ja muun liikenneinfran kanssa, ja kokonaisuus monitoroi paikallisia liikenneolojen, sään ja energiankulutuksensuhteita” (King, suomennos, Ollila 2014, 14–15).

Case 2. Teollisuus

”UPS optimoi reittejä ja säästää polttoainekuluissa autoihinsa kiinnitettyjen sensorien avulla” (Hammarsten 2013, 14). Teollisuusyritysten Big Datan käytöstä esimerkkeinä ovat General Electric, Schneider Electric ja Bosh, joista esimerkiksi General Electric keskittyy kehittämään koneista saatavan tiedon perusteella uusia palveluita ja tuotteita (Hammarsten 2013, 14).

Case 3. Terveydenhuolto

Big Data-työkalujen avulla on voitu maailmalla ennustaa influenssa-aaltojen leviämistä (Storås 2012,8). Big Datan analyysien soveltamisesta terveydenhuoltoon on malliesimerkkinä HealtMap.org, jossa palvelu louhii uutisia tarttuvista taudeista maailmanlaajuisesti eri mantereilla toimivista terveystietolähteistä sekä näyttää kartalta parhaillaan riehuvat epidemiat (Hämäläinen 2014, 17).

Case 4. Energiateollisuus

Oklahomalainen energiayhtiö OGE omistaa yhdeksän voimalaa. Yrityksellä on 758 000 asiakasta, joille yhtiön toimesta asennetuttiin älykkäät sähkömittarit. Uusien laitteiden avulla voidaan analysoida entistä tarkemmin sähkönkulutusta ja ennustaa kulutuspiikkejä. Uusi järjestelmä kerää dataa virrankulutuksesta kahden tunnin viiveellä. Aiemmin tämä tieto saatiin kuukausittain. (Pervilä 2012, 10.)

7 HAASTEET JA MAHDOLLISUUDET

Big Data luo yrityksille paljon uusia mahdollisuuksia, mutta tuo mukanaan myös haasteita. Yritysten kannalta Big Data voidaan nähdä uutena voimavarana, jonka kautta yrityksille syntyy uusia liiketoimintamahdollisuuksia. Tämä mahdollistaa aiempaa luotettavamman päätöksenteon. ”Suurista digitaalisista tietoaineistoista ja niiden hyödyntämisestä ennakoidaan uutta merkittävää tuotannon tekijää, joka mahdollistaa entistä parempaan tietoon perustuvan päätöksenteon niin yrityksille kuin julkisella sektorilla.” (Mikkelä 2012, 12.) Monenlaisille analytiikan osaajille löytyy tarvetta aina IT-asiantuntijoista data-analyytikoihin sekä henkilöihin jotka voivat yhdistää analytiikan liiketoimintaan ja viemään analyyttisia ratkaisuja osaksi liiketoimintaprosesseja (Krushe-Lehtonen 2014, 20).

7.1 Mahdollisuudet eri toimialoilla

Big Data luo mahdollisuuksia tuottavuuden parantamiseen ja kustannusten leikkaamiseen julkisella sektorilla. Se avulla voidaan tuottaa myös yrityksille aiempaa yksityiskohtaisempia ja tehokkaampia seurantaindikaattoreita. (Mikkelä 2012, 12.)

Terveydenhuoltoala

”Suomi on edelläkävijä terveydenhuollon tietojen sähköistämisessä ja sähköisten työkalujen käytössä. Käytännössä kaikki terveyskertomukset ovat jo sähköisiä ja potilastietoa kerätään kattavasti.” (Vänskä 2013, 8-9.)Terveydenhuollon alalla digitaalinen aineisto antaa terveyspalveluissa mahdollisuuden terveyden kehityssuuntien arvioimiseen esimerkiksi ehkäisemään epidemioita. Erilaisten hoitomuotojen tehokkuutta voidaan myös arvioida aineistojen avulla. (Melkas 2012, 53.)

Tutkimustiedon määrä lääketieteessä lisääntyy kiihtyvällä vauhdilla, jonka vuoksi sen analysointiin tarvitaan entistä kehittyneempiä menetelmiä (Hämäläinen 2014, 17.) ”Tutkittavat tietomäärät ovat niin suuria, että se asettaa reunaehtoja ohjelmistoille, koska esimerkiksi jo pelkästään yksi kokonainen auki sekvensoitu ihmisen genomi tuottaa dataa yli puoli teratavua ” (Hämäläinen 2014, 17). Terveydenhuollossa voitaisiin monilla tavoin hyödyntää suuria datamassoja, kun tietokoneella voitaisiin kertoa lääkärille vastaanotolle tulevan potilaan riskitekijöistä (Storås 2012, 8).

Analytiikkaosaamisella voitaisiin viedä sosiaali- ja terveydenhoitoalaa huomattavasti eteenpäin. Potilasturvallisuutta voitaisiin parantaa sekä vähentää hoitovirheitä yhdistämällä erilaisista tieto-lähteistä strukturoitua ja strukturoimatonta dataa. (Tuominen 2013, 27.)

Case Asiakaspalvelu

Big Datan avulla asiakaspalvelussa voidaan saada tarkempia asiakaskuvia ja asiakaspalvelun ammattilaiset voivat tehdä itsenäisempiä päätöksiä sen avulla. Yritysten BI-, CRM-, ERP- järjestelmiin ja tietokantoihin voidaan integroida verkon datavirtoja. Tämä mahdollistaa, että asiakkaista ja heidän käyttäytymisestään saadaan tarkka kuva. ”Yhdistämällä ennakoivat analytiikkaratkaisut ja käytettävissä olevat tiedot, yritykset voivat kehittää malleja, jotka ohjaavat organisaation ulkopuolelle suunnattuja toimintoja kuten asiakaspalvelua ja markkinointia” (Tarkempi asiakaskuva, 2012, 11).” Mobiililaitteiden tuottama data tarjoaa parhaimmillaan mahdollisuuden kehittää aivan uuttakin liiketoimintaa” (Hammarstein 2013, 14).

Case Vähittäiskauppa

Vähittäiskaupassa on huomattu Big Datan mahdollisuudet. Asiakasdatan analysointi antaa arvokasta, entistä tarkempaa tietoa asiakkaiden mieltymyksistä ja käyttäytymisestä. Yhdistämällä asiakasdataan ulkoista tietoa kuten säätiedot, voitaisiin ennakoida esimerkiksi lähipäivien juomien kysynnän kehittymistä. (Remes 2013, 9.)

Case Pankit

Pankit ovat keränneet ja seuloneet dataa jo vuosien ajan ja tehneet päätöksiä analyysien perusteella. Esimerkiksi muistinvaraisen analytiikan perusteella pankissa voitaisiin laskea kädenkäänteessä tilastollisia malleja, jotka eivät ole olleet aiemmin mahdollisia. (Storås 2012, 8.) Kertyvät data-aineistot tarjoavat erilaisia mahdollisuuksia pankeille ja rahoituslaitoksille. Niitä voidaan käyttää esimerkiksi rahoituksen riskien hallinnassa, auttamaan asiakkaille sopivien rahoitustapojen etsimisessä sekä eri toimialojen ja -alueiden kulutus- ja säästämistapojen profiloinnissa. Aineistot antavat näkemyksen esimerkiksi asiakkaiden maksuhistorioista, joiden avulla voidaan esimerkiksi arvioida luottoriskejä. (Melkas 2012, 52.)

7.2 Mahdolliset haasteet

Valtavalla nopeudella kertyvä, moninainen data tuo yrityksille uusia haasteita. Suurimmat haasteet tulevat liittymään datan käsittelyyn ja tallentamiseen. Useiden asiantuntijoiden mukaan yhtenä suurimmista Big Dataan liittyvistä haasteista pidetään pätevien analyytikkojen puutetta (Melkas 2012, 53). Haasteina tulevat myös olemaan tietoturvaan ja yksityisyyteen liittyvät asiat. Big Data- analyysien tullessa yrityksiin yhä tärkeämmäksi asiaksi tulee tietoturva ja datan suojaaminen. Tällaisia huomioitavia asioita ovat esimerkiksi terveydenhuoltoalalla potilastietojen ja henkilöiden yksityisyyden suojaaminen. Tämän kaltaiset turvallisuusvaatimukset ovat huomioitava jo sovelluksia tehdessä ja niissä tulee ottaa esimerkiksi huomioon, ketkä saavat nähdä potilastietoja ja missä olosuhteissa. (Hurwiz ym.2013, kappale 1)

Suuria tietomassoja ei voida käsitellä perinteisillä menetelmillä ja työkaluilla. Tiedon käsittelykapasiteetti ja välineet ja työkalut niiden käsittelemiseen tuovat oman haasteensa. Haasteena on myös, miten voidaan luoda ja kehittää teknologioita ja tietoarkkitehtuureja, joiden avulla mahdollistetaan taloudellisuus ja tehokkuus tiedon prosessointiin, louhintaan, analysointiin, visualisointiin ja varastointiin. (Mikkelä 2012, 12.)

Erilaiset laitteet ja anturit tuottavat informaatiota, jonka arvellaan kasvavan voimakkaasti tulevina vuosina. Suurenevat datamassat tuovat haasteita konesaleille. Suurin konesalien ongelma on verkkomuutoksiin kuluva aika. Ongelmia aiheuttavat palvelinten virtuaalisointi, suurten tietomassojen analysoinnin hitaus, tehon kulutus ja suorituskyky ja verkkomuutosten hidas toimitusaika-taulu. (Helenius 2013, 8.)

Kansainvälisenä ongelmana voidaan pitää data- ja analytiikkaosaajien vähäisyyttä. Maailmalla arvellaan Gartnerin ennustuksen mukaan olevan kysyntää 4,4 miljoonalle Big Data-työpaikalle vuoteen 2015 mennessä. Vuoden 2013 Tietojohtaja Ulla Krushe- Lehtosen mukaan Suomessa suurin tarve on Hadoop- ympäristössä työskennelleistä Big Data Developereista, Tietosuoja- ja -turvaosaajista sekä monipuolisista ratkaisuarkkitehdeista ja datavisualisoijista. Monenlaisille analytiikan osaajille

löytyy tarvetta aina IT-asiantuntijoista data-analyytikoihin sekä henkilöihin jotka voivat yhdistää analytiikan liiketoimintaan ja viedä analyyttisia ratkaisuja osaksi liiketoiminta-prosesseja (Krushe-Lehtonen 2014, 20)

8 YHTEENVETO

Big Data on ihmisen tai koneen synnyttämää strukturoitua tai strukturoimatonta dataa. Sitä syntyy jatkuvasti kiihtyvällä nopeudella, monenmuotoisena ja kokoisena erilaisista lähteistä. Se on muodoltaan niin suurta ja vaihtelevaa, ettei sitä voida käsitellä perinteisin tiedonhallintamenetelmin. Erilaisia tiedonlähteitä ovat esimerkiksi sosiaalisen median alustat, yritysten erilaiset sovellukset ja järjestelmät. Tietoa tuottavat myös erilaiset anturit ja sensorit.

Dataa on paljon ja se sisältää strukturoimatonta ja strukturoitua dataa. Sitä ei voida hallita, käsitellä ja varastoida perinteisin tiedonhallintamenetelmin. Suuren ja moninaisen datan louhintaan, varastointiin, analysointiin ja jalostamiseen tarvitaan uudenlaisia työkaluja. Tällaisen vaihtoehdon tarjoaa Hadoop avoimen lähdekoodin ohjelmistokehys oheisprojekteineen. Hadoop on NoSQL- tietokanta, joka tarjoaa ratkaisun strukturoimattoman datan varastointiin ja -hallintaan. Se sisältää useita työkaluja, jotka ovat suunniteltu toimimaan yhdessä Hadoop- ympäristössä. Työvälineiden avulla data saadaan tallennettua ja muutettua sellaiseen muotoon että sitä voidaan hyödyntää liiketoiminnan tarpeisiin.

Tietoturva ja yksityisyys liittyvät voimakkaasti Big Data-ilmiöön. Suomen laissa on säädetty yksityisyyttä ja henkilötietoja koskevasta lainsäädännöstä. Big Data- sovellusten yleistyessä on mahdollista, että tätä lakia tullaan rikkomaan. Ihmisistä kerätään paljon tietoja erilaisiin järjestelmiin, ja riskinä on myös, että näistä saatujen tietojen perusteella tiedon haltijalla saattaa olla erittäin paljon tietoa henkilöstä. Terveydenhuollossa potilastietoja kertyy paljon, ja ne luokitellaan arkaluontoisiksi tiedoiksi. Myös valokuvat, äänitiedostot ja videot, joista henkilö on tunnistettavissa, voidaan tulkita henkilötiedoiksi.

Pilvipalvelut liittyvät voimakkaasti Big Data- ilmiöön. Pilvipalvelu mahdollistavat yrityksille halpaa tallennustilaa ja kustannussäästöjä. Pilvipalvelut mahdollistavat myös Big Datan käytön pienille yrityksille. Yritysten sähköistäessä toimintaansa ja siirtäessään palvelujaan pilveen tullaan myös kohtaamaan tietoturva ja yksityisyyden suojaan liittyviin huoliin. Big Dataa käytetään muualla maailmassa monipuolisesti eri aloilla. Suomi on kehityksestä jäljessä vuosia verrattuna muuhun maailmaan, myös pilvipalveluiden osalta.

9 POHDINTA

Opinnäytetyön tavoitteena oli tehdä tietopaketti siitä mitä Big Data on, missä sitä syntyy ja mitkä ovat sen lähteet. Työn alussa selvitettiin Big Datan sisältämät datatyypit ja avattiin hieman käsitteitä tiedonlouhinta, tiedon varastointi, tiedon jalostaminen ja tiedon analysointi. Big Dataan liittyvistä työkaluista ja menetelmistä tarkasteltiin pintapuolisesti avoimen lähdekoodin Hadoop- sovellusta ja siihen liittyviä muita oheisprojekteja. Tietoturva ja yksityisyys -kappaleessa avattiin hieman, mitä ne ovat ja miten asia liittyy Big Data -ilmiöön. Tietoturvakohdan ja yksityisyydensuoja -asian lisäsin työhöni, koska nämä ovat sellaisia asioita ja mahdollisia haasteita / ongelmia, joihin tullaan Big data - ilmiössä törmäämään.

Pilvipalvelu kappaleessa kerrottiin mitä palvelut ovat ja miten liittyvät asiaan. Työssä käytiin läpi muutamien lyhyiden case esimerkkien kautta miten ja missä Big Dataa käytetään Suomessa ja muualla maailmassa. Tämän työn tietoperusta perustuu olemassa olevaan kirjallisuuteen, elektronisiin julkaisuihin ja lehtiartikkeleihin.

Työn tekeminen oli haastavaa ja työlästä, koska alkutietoa aiheesta minulla ei ollut. Myös tiedonlouhinta, tiedon analysointi, tiedon varastointi ja tiedon jalostaminen olivat uusia asioita, joita ei koulussamme ollut opiskeltu ja ne vaativat asiaan perehtymistä. Suomenkielistä kirjallisuutta Big Datasta ei aloittaessani työtä vielä löytynyt kuin yksi ja Hadoopista löytyi vain yksi gradu. Lehtiartikkeleita löytyi jonkin verran, mutta mielestäni niiden tietosisältö oli vähäistä. Työn alussa tiedonhakuun meni runsaasti aikaa. Lopulta englanninkielistä kirjallisuutta aiheesta löytyi runsaasti, mutta englanninkielisten aineistojen läpikäyminen oikeanlaisen tiedon löytämiseksi oli mielestäni haastavaa. Kuvailisinkin opinnäytetyöprosessiani samankaltaiseksi kuin itse aihe. Suurista tietomääristä oli löydettävä ja louhittava sellaista tietoa, jota voitaisiin käyttää työssäni. Tietoa oli paljon ja sitä oli aluksi vaikea jäsennellä ja yrittää hahmottaa asioiden yhteyksiä.

Palkitsevana koin kuitenkin sen tunteen kun, asia alkoi avautua ja pikkuhiljaa kertynyt tieto alkoi muuttua ymmärrykseksi. Haasteena koin kuitenkin, kuinka kertoisin sen työssäni hyvin yksinkertai-

sesti ja ymmärrettävästi muillekin. Siksi päätin lopulta aloittaa työni määrittelemällä lyhyesti mitä on data sekä mitkä datatyypit ovat ominaista Big Data -käsitteelle.

Aihe oli laajuudeltaan vieläkin suurempi kuin miltä se aluksi vaikutti. Aiheen laajuuden vuoksi työ oli rajattava aiottua suppeammaksi eikä asioita voitu käydä kovin syvällisesti läpi. Monia uusia kysymyksiä heräsi aiheesta ja toivoisinkin, että joku kiinnostuisi aiheesta ja jatkaisi tästä eteenpäin. Osa-alueita on paljon ja uskoisin sieltä löytyvän jokaiselle suuntautumisvaihtoehdollekin omanlaisensa alue. Erityisen tärkeänä pitäisin että aihetta tarkasteltaisiin tietoturvan ja yksityisyyden näkökannoista. Myös työvälineitä ja käytettäviä menetelmiä voitaisiin tarkastella yksityiskohtaisemmin. Lisää tietoa voitaisiin hankkia NoSQL- tietokannoista ja niiden eri tyyppejä voitaisiin tutkia tarkemmin. Tiedon visualisointi on myös aihealue, joka on rajattu tästä työstä pois. Sitä voitaisiin myös tarkastella aiheeseen liittyen. Big Datan antamia mahdollisuuksia ja uhkia julkisen puolen organisaatioille, kuten terveydenhuollon alalle voitaisiin myös tutkia.

Tämän opinnäytetyön tuloksena syntyi pienimuotoinen tietopaketti, jossa on käsitelty Big Dataa aina datasta, sen lähteistä, tallentamisesta, työmenetelmistä ja tietoturvasta käyttöön asti. Tietopaketin avulla voidaan tutustua Big Data aiheeseen.

LÄHTEET

Alanko, M & Salo, I. 2013. Big data suomessa keskustelualoite, 25/2013, Hakupäivä 28.2.2014, http://www.lvm.fi/docs/fi/2497123_DLFE-21601.pdf

Apache.org.2014.http://hbase.apache.org/index.html, hakupäivä 2.6.2014

Andersson, A, Koivisto, J & Ylipartanen, A.2013. Tietosuojavastaavan käsikirja. Helsinki: Tietosanoma Oy.

Capriolo, E, Wampler, D & Rutherglen, J. 2012. Programming Hive, Data Warehouse and Query LaguageforHdoop.O'ReillyMedia,Hakupäivä14.2.2014,Http://proquest.safaribooksonline.com/book/databases/hadoop/9781449326944/1dotintroduction/_an_overview_of_hadoop_and_mapreduce_html?uicode=ouluuas

Gates, A. 2011. Programming Pig. O'Reilly: Sebastobol: CA, Hakupäivä21.2.2014,Http://cdn.oreilly.com/oreilly/booksamplers/9781449302641_sampler.pdf

Helenius, T.2013. Paisuvat datamassat konesalien haasteena. Tietokone 9, 8

Henkilötietolaki 22.4.1999/523. www.finlex.fi

Hovi, A, Hervonen, H, Koistinen, H.2009. Tietovarastot ja business intelligence. WSOY: Porvoo, 14

Hammarsten H. 2013. Big data jyllää teollisuudessa. Tietoviikko 15.11.2013, 14.

Hotti, M. 2012. Pikaperehdytys Big Dataan, Hakupäivä 12.5.2014, http://www.tietoviikko.fi/msareena/msteemat/bi/pikaperehdytys+big+dataan+mika+on+apache+hadoop+enta+hive/a859947

Hurwiz, J, Nugent, A, Halper, F & Kaufman, M.2013. Big Data For Dummies. Canada: John Wiley & Sons, Inc. Hakupäivä 11.2.2014,Http://proquest.safaribooksonline.com/book/databases/business-igence/9781118644171/chapter2examiningbigdatatypes/a2_06_9781118644171_ch02_html?uicode=ouluuas

Hämäläinen P. 2014. Tehoa terveydenhuoltoon. Tietokone 2, 17

Hämäläinen, P.2012. Big Data jalostaa tiedon rahaksi. Tietokone 12, 36 -37.

Jain, A. 2013. Instant Apache Sqoop. Packt Publishing: Birmingham B3 2PB, UK. Hakupäivä 19.2.2014,Http://proquest.safaribooksonline.com/book/operating-systems-and-server-administration/apache/9781782165767/1dot-instant-apache-sqoop/ch01s05_html?uicode=ouluuas

Junqueira, F & Reed, B. 2013. ZooKeeper. O'Reilly Media, Inc. Hakupäivä 17.2.2014, Http://proquest.safaribooksonline.com/book/operating-systems-and-server-administration/apache/9781449361297/idot-zookeeper-concepts-and-basics/ch01_html?uicode=ouluuas

Järn, S. 2013. Uutista mä metsästän. Tietoasiantuntija.1, 9-10

Järvinen. P. 2010. Yksityisyys, turvaa digitaalinen kotirauhasi. Jyväskylä: WSOY pro Oy.

Järvinen. P. 2012. Arjen tietoturva. Jyväskylä: WSOY pro Oy.

Karttunen, T, Laasanen, H, Sippel, L, Uitto, T & Valtonen, M. 2012. Juridiikan perusteet. Helsinki: Sanoma Pro Oy.

King, J käännös Ollila, K.2014. Analytiikka pelasti Fordin. Tietoviikko 1, 14 -15.

Kiiski, M. 2013. IT myllertää terveydenhuollon. Tietoviikko 20.9.2013, 10

Krishnan, K. 2013. Data Warehousing in the Age of Big Data. Hakupäivä 10.2.2014, Morgan Kaufmann:USA.Http://proquest.safaribooksonline.com/book/databases/data-warehouses/9780124058910/chapter-2dot-working-with-big-data/st0025_chp002_html?uicode=ouluuas

Krushe - Lehtonen, U. 2014. Mistä sopivia tietoasiantuntijoita? Tietoasiantuntija 1, 20.

Liikenne- ja viestintäministeriö. Alanko, M, Salo, I. Big data Suomessa - keskustelualoite, 25/2013 s.7-8. Hakupäivä 28.2.2014, Http://www.lvm.fi/docs/fi/2497123_DLFE-21601.pdf

Lublinsky, B, Smith, K, Yakubovich, A. 2013.Professional Hadoop Solutions. Hakupäivä 18.3.2014,Http://proquest.safaribooksonline.com/book/databases/hadoop/9781118824184/chapter-1-big-data-and-the-hadoop-ecosystem/chapter01_html?uicode=ouluuas

Lång, J. 2014. Big Data, henkilötiedot ja yksityisyyden suoja-kehittyvä teknologia ja lainsäädäntö törmäyskurssilla? Hakupäivä 6.5.2014, Http://www.bigdata.fi/blogi/vierailijakirjoitus/big-data-henkilotiedot-ja-yksityisyyden-suoja-kehittyva-teknologia-ja

Manninen, I. 2014. Big data hankkeet-kasvua onnistumisen kautta. Hakupäivä 7.5.2014,Http://www.bigdata.fi/blogi/vierailijakirjoitus/big-data-hankkeet-kasvua-onnistumisien-kautta

Melkas, J.2012. Iso data – suuret lupaukset ja pullonkaulat. Tieto & trendit 4-5, 52 -53.

Microsoft Azure.2014. Hakupäivä 12.6.2014, http://azure.microsoft.com/en-us/services/hdinsight/

Mikkelä, H.2012. Big Data mullistaa tietomaailmaa? Tietoasiantuntija 4, 11–12.

Patil, M & Thia, F. 2013 Pentaho for Big Data Analytics. Packt Publishing: Hakupäivä 25.2.2014, Http://proquest.safaribooksonline.com/book/programming/java/9781783282159/3dot-churning-big-data-with-pentaho/ch03s02_html?uicode=ouluuas

Prajapati, V. 2013. Big Data Analytics with R and Hadoop. Birgingham B3 2PB, UK: Packt Publishing. Hakupäivä20.3.2014,Http://proquest.safaribooksonline.com/book/databases/hadoop/9781782163282/preface/pr07s04_html?uicode=ouluuas

Pentaho for Big Data Analytics. Hakupäivä 25.2.2014, Http://proquest.safaribooksonline.com/book/programming/java/9781783282159/1dot-the-rise-of-pentaho-analytics-along-with-big-data/ch01_html?uicode=ouluuas

Pervilä, M. 2012. Missä on Big datan hyöty. Tietoviikko 5.11.2012, 10

Pitkänen, O, Tiilikka, P & Warma, E. 2013. Henkilötietojen suoja. Vantaa: Talentum.

Remes, M. 2013. Big datasta kaikki hyöty irti. Hakupäivä 6.5.2014, Http://issuu.com/kpmgfinland/docs/asiakaslehti-view-1-2013/8?e=6625711/1463933

Ruckenstein, M. 2013. Hallitseva ja houkuttava big data-tieteen ja teknologiantutkimuksen aihioita. Tieteessä tapahtuu.6, 34.

Salo, I. 2013. Big Data, tiedon vallankumous. Jyväskylä: Docento.

Siiramaa, P. 2012. HADOOP Ohjelmistokehitys Big Datan käsittelyyn. Itä-Suomen Yliopisto. Luonnontieteiden ja metsätieteiden tiedekunta. Pro- gradu-tutkielma

Schmidt, K & Phillips, C. 2013. Programming Elastic MapReduce. Canada: O'Reilly Media, Inc. Hakupäivä12.2.2014,Http://proquest.safaribooksonline.com/book/databases/hadoop/9781449364038/2dot-data-collection-and-data-analysis-with-aws/_understanding_mapreduce_html?uicode=ouluuas

Siltala, T.2014. Big data tuli busseihin, Tietoviikko 1/2014,

Storås, N. 2012. Suuri on mutkikasta – Suomi heräsi big data aikaan liian myöhään. Uusien työkalujen tarjoamat hyödyt uhkaavat vesittyä. Tietoviikko 5.11.2012, 8

Suomen Internetopas. Tietoturva. Hakupäivä 25.3.2014, http://www.internetopas.com/yleistietoa/tietoturva/)

Tietosuojavaltuutetun toimisto, Rekisterinpitäjälle, 5. Käyttötarkoituksen määrittely ja käsittelyn suunnittelu. Hakupäivä 18.3.2014, http://www.tietosuoja.fi/1698.htm

Tuominen, L. 2013. Big Data- uhka vai mahdollisuus? Tietoasiantuntija.1, 2013

Vaish, G. 2013. Getting Stardet with NoSQL. Packt Publishing: Birmingham B3 2PB, UK. Hakupäivä 19.2.2014, Http://proquest.safaribooksonline.com/book/-/9781849694988/1dot-an-overview-of-nosql/ch01s02_html?uicode=ouluuas

Valtiovarainministeriö. Henkilöstön tietoturva 4/2013. Hakupäivä 25.3.2014, Http://www.vm.fi/vm/fi/04_julkaisut_ja_asiakirjat/01_julkaisut/05_valtionhallinnon_tietoturvallisuus/20131122Henkil/Vahti_4_2013_A5.pdf

Vuokola, J. 2013. Puhu minulle datasta. Tietoviikko 20.9.2013, 9.

Väestörekisterikeskus. Henkilötunnus. Hakupäivä 25.3.2014, Http://www.vrk.fi/default.aspx?id=167

Vänskä, O. 2013. IT myllertää terveydenhuollon. Tietoviikko 20.9.2013, 8 -10.

White, T. 2012. Hadoop: The Definetive Guide, 3rd Edition. CA: O' Reilly Media, Inc.

White, T. 2012.Hadoop: the define guide Sebastopol CA: O'Reilly. Hakupäivä.13.2.2014 http://proquest.safaribooksonline.com/book/software-engineering-and-development/9781449328917/2dot-mapreduce/id2407783?uicode=ouluuas

Zburivsky, D.2013.Hadoop Cluster Deployment. Birmingham B 3 2PB UK: Packt Publishing. Hakupäivä17.2.2014,Http://proquest.safaribooksonline.com/book/databases/hadoop/9781783281718/3dot-configuring-the-hadoop-ecosystem/ch03s02_html?uicode=ouluuas

LIITTEET

LYHENNELUETTELO

Algoritmi	Yksityiskohtainen kuvaus tai ohje prosessin suorittamiseen, jonka avulla ongelma voidaan ratkaista
Big Table	Googlen kehittämä hajautettu varastointijärjestelmä strukturoidun datan hallintaan
Klusteri	Useista tietokoneista muodostuva verkotettu malli, hajautettuun tiedonkäsittelyyn ja laskentaan
HDFS	Hadoop Distributed File System. Hajautettu tiedostojärjestelmä, jota käytetään Hadoopissa
GFS	Google File System, Googlen kehittämä hajautettu tiedostojärjestelmä massiivisten tietomäärien tallentamiseen
NFS	Network File System (NFS) on hajautettu tiedostojärjestelmä
NoSQL	NoSQL (Not only SQL) on tietokantamalli, joka perustuu avain/arvopareihin. Tallennettavaa dataa ei tallenneta tauluihin tai sarakkeisiin toisin kuin perinteisemmissä relaatiotietokannoissa.
SQL	Structured Query Language on melkein kaikkien relaatiotietokantatuotteiden tukema tietokannan määrittely- ja käsittelykieli
Relaatiotietokanta	Taulujen välille luodaan yhteyksiä. Taulut yhdistetään toisiinsa toisen taulun avaimella.
RDBMS	Relaatiotietokantojen hallintajärjestelmä
Tietokanta	Kokoelma tietoja, joilla on yhteys toisiinsa