Mika Sutinen

BIG DATA JA ANALYTIIKKA TERVEYDENHUOLLOSSA

JYVÄSKYLÄN YLIOPISTO
TIETOJENKÄSITTELYTIETEIDEN LAITOS
2016

TIIVISTELMÄ

Sutinen, Mika
Big Data ja Analytiikka terveydenhuollossa
Jyväskylä: Jyväskylän yliopisto, 2016, 36 s.
Tietojärjestelmätiede, kandidaatin tutkielma
Ohjaaja: Seppänen, Ville

Elämme datakeskeisessä maailmassa, jossa uutta digitaalista sisältöä syntyy jo muutamissa sekunneissa uskomattomia määriä. Yritykset ja organisaatiot pyrkivät analysoimaan tätä massiivista, monimuotoista ja nopeasti kasvavaa big dataa toimiakseen paremmin ja tehokkaammin sekä saadakseen parempaa ymmärrystä omasta toiminnastaan. Big data ja analytiikka ovat aihe, jota on tutkittu laajasti myös terveydenhuollon näkökulmasta. Terveydenhuolto on yksi toimialoista, joka voi hyötyä big datasta ja siihen kohdistuvasta analytiikasta huomattavasti. Tässä tutkielmassa kysytään kuinka analytiikka on kehittynyt ja mihin suuntaan se on kehittymässä. Kysymykseen vastataan tutustumalla liiketoimintatiedon hallinnan ja analytiikan historiaan ja kehitykseen sekä teknologioihin, jotka mahdollistavat big data -analytiikan. Tutkielman toisena kysymyksenä on kuinka analytiikkaa voidaan hyödyntää terveydenhuollossa. Tähän vastataan esittelemällä erilaisia analyyttisia ratkaisuja, joita terveydenhuollot organisaatiot maailmalla ovat toteuttaneet. Tämän lisäksi tutkielmassa tehdään lyhyt katsaus siihen millaisia analytiikkaan pohjautuvia ratkaisuja olisi mahdollista toteuttaa Suomessa. Tutkielmassa todetaan analytiikan kehityksen painottuvan big datan analytiikkaan ja terveydenhuollon toimialalla olevan lukuisia eri mahdollisuuksia analytiikan käyttöön. Tutkielma on toteutettu kirjallisuuskatsauksena käyttäen pääsääntöisesti luotettavaksi todettuja tieteellisiä julkaisuja. Joissain tapauksissa viitataan myös kaupallisten yritysten tuottamiin teknisiin kuvauksiin (*white paper*) ja organisaatioiden tai yritysten verkkosivuihin. Näitä viittauksia käytetään kuitenkin vain lisätietojen saamista varten spesifeistä aiheista, ei varsinaisina tieteellisen tutkimuksen lähteinä.

Asiasanat: Analytiikka, big data, koneoppiminen, terveydenhuolto, liiketoimintatiedon hallinta, Web 2.0, Esineiden Internet

ABSTRACT

Sutinen, Mika
Big Data and Analytics in Healthcare
Jyväskylä: University of Jyväskylä, 2016, 36 p.
Information Systems, Bachelor's thesis
Supervisor: Seppänen, Ville

We live in a data-centric world where new digital content is created in incredible amounts in matter of seconds. Companies and organizations attempt to analyze this massive, diverse and fast growing big data to function better and more effectively while also looking to gain new insights on how they operate. Big data and analytics are subjects that have been researched thoroughly from the healthcare point of view. Healthcare is also one of the industries that is widely recognized to be among the top beneficiaries of big data analytics. In this thesis we're looking for answers to question on how analytics have developed and to what direction analytics are being developed. This question is answered by examining the history of business intelligence & analytics, the development and technologies that make big data analytics possible. The second question posed in the thesis is how analytics can be used in healthcare industry. The answer is provided by doing review of analytics solutions implemented by healthcare organizations around the world. The thesis also includes a short review on what kind of analytic solutions could be done in the Finnish healthcare. In this thesis the development of analytics is discovered to be big data oriented and that there are many opportunities in healthcare industry for using advanced analytics. This thesis has been done as a literature review, the sources coming from well-known and reliable scientific publications. There are some references to white papers by information technology companies and to company or organizational web-pages. They are only used for providing information on specific topics and are not considered to be main sources for scientific research.

Keywords: Analytics, big data, machine learning, healthcare, business intelligence, Web 2.0, Internet of Things

KUVIOT

TAULUKOT

SISÄLLYS

1 DATA, UUSI LUONNONVARA

Datasta puhutaan tänä päivänä usein uutena luonnonvarana ja tämä näkemys on helposti hyväksyttävissä, sillä elämme maailmassa, joka on erittäin datakeskeinen. Siitä hetkestä, kun aloitit tämän kappaleen lukemisen, digitaalisen maailman sisältö on kasvanut uskomattomalla nopeudella. Noin 10 sekunnin aikana Instagramiin on lisätty 300 000 valokuvaa, Twitteriin 60 000 uutta twiittiä, Facebookiin 700 000 tykkäystä ja YouTubeen 50 tuntia videota (katso liite 1). Luonnonvaroille tyypillistä on myös se, että niitä pyritään hyödyntämään taloudellisen edun saamiseksi. Tässäkään data ei muodosta poikkeusta. Vuonna 2011 Bloomberg Businessweekin suorittamassa katsauksessa liiketoiminnan analytiikan käyttöön tulokset kertoivat, että yrityksistä joiden liikevaihto ylitti 100 miljoonaa, 97 % käytti analytiikkaa jossain muodossa (Chen, Chiang & Storey, 2012).

Yritysten tallentaman datan hyödyntäminen ei ole uusi asia vaan sen juuret löytyvät 90-luvulta. Silloin syntyivät ensimmäiset yleisesti saatavat analytiikan mahdollistavat teknologiat ja informaatioteknologian määritelmä business intelligence (*BI*), suomeksi liiketoimintatiedon hallinta, vakiintui käyttöön. 2000-luvulla analytiikassa kiinnostus siirtyi aikaisemmin tehtyjen toimenpiteiden vaikutusten raportoinnista tulevien tapahtumien ennustamiseen. Tämän kehityksen taustalla suurimpia vaikuttajia olivat Web 2.0 teknologiat, jotka mahdollistivat nykyisen kaltaiset sosiaalisen median palvelut ja loivat suunnattomia määriä analysoitavaa dataa. Tämän, monimuotoisen ja usein strukturoimattoman, datan analysointiin ja muuttamiseksi liiketoimintaa edistäväksi ymmärrykseksi tarvittiin uusia teknologioita ja toimintatapoja (Chen ym., 2012). Näin vakiintunut business intelligence määritelmä kehittyi kattamaan laajemmin analytiikan ja syntyi käsite business intelligence & analytics (BI&A).

Tänä päivänä elämme big datan aikakautta. Viimeiset vuodet maailmalla vallalla ollut voimakas big data hype ei näytä merkkejä laantumisesta, päinvastoin. Eräiden arvioiden mukaan maailmassa luodun datan määrä jatkaa kasvamistaan tuplaantuen kahden vuoden välein ainakin vuoteen 2020 saakka. Tämä arvio saattaa osoittautua jopa hyvin varovaiseksi Internet of Thingsin (IoT) yleistyessä ja yhä useampien laitteiden kytkeytyessä tietoverkkoihin. Useat alan

toimijat ja analyytikot, mm. Ericsson, Microsoft, Gartner ja McKinsey Global Institute, ovat arvioineet vuonna 2020 IoT laitteita olevan 20–30 miljardia kappaletta. Kun huomioidaan datamäärien nopea kasvu ja edellä esitetyt arviot, on selvää että myös analytiikassa käytettyjen teknologioiden ja toimintatapojen tulee kehittyä nopeasti. Big data tuo mukanaan myös uudenlaisia haasteita sille kuinka ja missä analysointi tapahtuu. Vaikka tallennusteknologiat ovat kehittyneet niiden nopeuden ja kapasiteetin kasvaessa, nykyisin kaikkea dataa ei voida enää tallentaa pysyvästi, vaan datan analysointi tapahtuu yhä useammin sen ollessa vielä liikkeessä.

Big data ja analytiikka ovat myös terveydenhuollossa merkittävässä roolissa. Yhä useammin terveydenhuollon järjestelmistä syntyvä data on strukturoimatonta ja sisällöltään muuta kuin tekstiä, kuten videota, ääntä, lääketieteellisiä kuvia tai monitorointilaitteiden tietovirtoja. Monimuotoisuuden lisäksi dataa syntyy terveydenhuollossa paljon. Potilastietojärjestelmien lisäksi dataa on saatavissa erilaisista monitorointilaitteista ja sensoreista, sosiaalisen median lähteistä sekä resepti- ja päätöksentuenjärjestelmistä. Terveydenhuollossa tärkeässä roolissa ovat myös datan käsittelyn nopeus, sillä potilaiden terveys tai selviytyminen ovat riippuvaisia oikea-aikaisesta hoidon aloituksesta tai riskien tunnistamisesta. Kehittyneellä analytiikalla on terveydenhuollossa useita potentiaalisia käyttökohteita; sairauksia voidaan tunnistaa nopeammin, väestön terveydentilan muutoksia voidaan seurata tarkemmin sekä saadaan hyödyllistä tietoa kliinisesti parhaista ja kustannustehokkaista hoitomuodoista. (Raghupathi & Raghupathi, 2014)

Tässä tutkielmassa perehdytään datan analysointiin, analytiikan kehitykseen ja sen hyödyntämiseen terveydenhuollossa. Terveydenhuolto on yksi niistä toimialoista, joka useissa tutkimuksissa mainitaan yhtenä suurimpina hyötyjistä big data analytiikassa (Kambatla, Kollias, Kumar & Grama, 2014). Kotimaisesta näkökulmasta katsottuna aihe on myös mielenkiintoinen, sillä Suomessa sähköisiä potilastietojärjestelmiä on käytetty jo kauan. Analytiikka terveydenhuollossa ei lähtökohtaisesti tarkoita vain kustannuksissa säästämistä, vaikka tämä toteutuu monissa esimerkkitapauksissa. Yhtä tärkeänä, tai ehkä tärkeämpänäkin tavoitteena, nähdään hoidon ja elämisen laadun parantuminen terveydenhuollon organisaatioiden pystyessä palvelemaan asiakkaitaan nopeammin sekä auttamaan tehokkaammin vaivojen ja sairauksien ennaltaehkäisyssä. Tutkielmassa haetaan vastauksia seuraaviin tutkimuskysymyksiin:

- Kuinka analytiikka on kehittynyt ja mihin suuntaan se on kehittymässä?
- Kuinka analytiikkaa on hyödynnetty terveydenhuollon organisaatioissa maailmalla ja kuinka sitä voidaan hyödyntää suomalaisessa terveydenhuollossa?

Tutkielman toteutustapa on kandidaatin tutkielmalle tyypillinen kirjallisuuskatsaus. Lähdemateriaalia tutkielmaan on etsitty yliopiston käytössä olevista informaatioteknologian tiedekannoista kuten ProQuest ja IEEE sekä Google Scholar hakukoneella. Kaupallisten toimijoiden ja yritysten, kuten IBM ja Mic-

rosoft, ratkaisukuvausten yhteydessä lähteenä käytetään näiden omia materiaaleja. Hakukoneissa materiaalin etsimiseen käytettiin mm. seuraavia asiasanoja: *healthcare, analytics, big data, cep, complex event processing, mapreduce, hadoop, medical imaging, stream analytics.* Koska kehittynyt analytiikka terveydenhuollossa on melko tuore aihe, rajoitettiin joissain tapauksissa tulokset koskemaan vain vuoden 2010 ja sen jälkeen tehtyä tutkimusta.

Tutkielman rakenne on seuraava. Johdantoa seuraavassa luvussa, joka on ensimmäinen sisältöluku, käydään läpi liiketoimintatiedon hallinnan ja analytiikan kehitystä sekä esitellään analytiikan osalta teknologioita ja käsitteitä. Kolmannessa luvussa esitellään terveydenhuollon kannalta tärkeiksi tunnistettuja analytiikan trendejä sekä esitellään terveydenhuollon analyyttisia ratkaisuja, joita maailmalla on kehitetty. Luvun lopussa esitetään ajatuksia mahdollisista terveydenhuollon analyyttisistä ratkaisuista Suomessa. Neljäs ja viimeinen luku sisältävät tutkimustulosten yhteenvedon sisältölukujen osalta sekä pohdinnan mahdollisista jatkotutkimusaiheista.

2 LIIKETOIMINTATIEDON HALLINTA JA ANALYTIIKKA (BI&A)

Tässä luvussa tarkastellaan liiketoimintatiedon hallinnan ja analytiikan kehitystä. Luvun rakenne on seuraava. Ensimmäisenä käydään läpi liiketoimintatiedon hallinnan ja analytiikan kehitystä yleisellä tasolla sekä esitellään yleisesti tunnistettuja kehityssuuntia ja näihin liittyviä haasteita, teknologioita ja uusia paradigmoja. Seuraavaksi käsitellään liiketoimintatiedon hallinnan ja analytiikan sekä tämän tutkielman kannalta keskeisiä käsitteitä. Luvun lopussa käydään läpi tarkemmin liiketoimintatiedon hallinnan ja analytiikan historiaa sekä esitetään arvioita tulevasta kehityksestä. Luvussa vastataan siis tutkielmassa esitettyyn kysymykseen siitä, kuinka liiketoimintatiedon hallinta ja analytiikka on kehittynyt ja mihin suuntaan se on kehittymässä.

Analytiikan kehityksen kannalta merkittävässä roolissa on ollut tietokoneiden komponenttien, kuten prosessorien, tallennusjärjestelmien ja muistien, kehitys. Tietokoneiden suorituskyvyn ja tallennuskapasiteetin kasvaminen on mahdollistanut yhä suurempien datamäärien tallentamisen ja analysoinnin kustannustehokkaasti. Myös tietokoneiden keskusmuistien määrät ovat kasvaneet komponenttien hintojen pysyessä kohtuullisina. Tämän ansiosta yksittäisellä tietokoneella voi tänä päivänä olla satoja tai tuhansia gigatavuja keskusmuistia. Tämä kehitys on edistänyt myös äärimmäisen suorituskykyisten muistinvaraisten (in-memory) analyyttisten järjestelmien kehitystä. (Chaudhuri, Dayal & Narasayya, 2011.)

Analytiikassa käytettävät datamäärät kasvavat kuitenkin huomattavasti nopeammassa tahdissa kuin mitä tietokoneiden suorituskyky on kehittynyt. Esimerkiksi vuosien 2002–2009 välisenä aikana dataliikenteen määrä kasvoi 56-kertaiseksi tietokoneiden suorituskyvyn kasvaessa noin 16-kertaiseksi. Tätä epätasapainoa tasaamaan on syntynyt useita teknologioita, joista yksi tärkeimmistä on virtualisointi, joka muodostaa modernien pilvipalveluiden selkärangan. Virtualisointiteknologiat mahdollistavat suorituskykyiset ja vikasietoiset palvelinkeskusarkkitehtuurit, joissa yksittäisistä palvelimista on voitu rakentaa laajoja analytiikka-klustereita. (Kambatla ym., 2014)

Pilvipalvelujen suosiota analytiikassa on edistänyt palvelusuuntainen ajattelu, jota voidaan pitää yhtenä viime vuosien IT-alan voimakkaimmin kasvaneista paradigmoista. Sellaiset pilvipalvelut, jotka tarjoavat laskenta- ja tallennuskapasiteettia käyttöön perustuvalla hinnoittelumallilla, ovat tuoneet suurta kapasiteettia vaativan analytiikan myös pienien ja keskisuurten yritysten ja organisaatioiden saataville tarjoamalla siihen kustannustehokkaan tavan. (Demirkan & Delen, 2013.) Vaikka erilaisia pilvipalveluja on ollut saatavilla jo useita vuosia, analytiikka palveluna (Analytics-as-a-Service), on suhteellisen uusi nousija pilvipalveluiden listalle. Tähän syiksi on mainittu mm. analyyttisten mallien ylläpidon ja hallinnan monimutkaisuus, analytiikassa tarvittavan teknologian kehitys sekä standardointi. (Delen & Demirkan, 2013.)

Web 2.0 ja siihen liittyvät teknologiat ja palvelut loivat tarpeen analysoida suuria määriä nopeasti syntyvää ja monimuotoista käyttäjien luomaa sisältöä. Ne aloittivat big data analytiikan aikakauden. Seuraavassa kehitysvaiheessa kymmenet miljardit verkkoon liitetyt sensorit, rakennukset, kodinkoneet ja nopeasti yleistyvät mobiililaitteet, jotka muodostavat esineiden Internetin, Internet of Thingsin, tulevat vaatimaan uudenlaista ajattelua ja teknologioita analytiikan saralla. (Chen ym., 2012; Fan & Bifet, 2013.)

2.1 Liiketoimintatiedon hallinnan ja analytiikan käsitteitä

Liiketoimintatiedon hallinta ja analytiikka on laaja kokonaisuus, joka pitää sisällään suuren määrän erilaisia teknologioita, toimintatapoja ja paradigmoja joiden kaikkien käsittely kandidaatin tutkielman yhteydessä ei olisi mielekästä, eikä edes mahdollista. Seuraavaksi esitellään lyhyesti ne käsitteet, jotka ovat tämän tutkielman ymmärtämisen kannalta oleellisia. Esiteltävien käsitteiden valintaan on vaikuttanut myös se, kuinka usein ne esiintyivät tutkielmassa käytetyissä lähdemateriaaleissa ja kuinka tärkeäksi niiden rooli on sitä kautta tunnistettu analytiikkaan liittyvässä tutkimuksessa.

Analytiikka

Analytiikalla tarkoitetaan toimintaa, jossa käytetään erilaisia tilastollisia menetelmiä olemassa olevaan dataan. Tavoitteena on kehittää yrityksen tai organisaation toimintaa vastaamalla esimerkiksi seuraaviin kysymyksiin; Mitä on tapahtunut, miksi jotain tapahtuu ja mitä pitää tehdä, että haluttu tapahtuma saadaan aikaiseksi. Monet analytiikassa käytetyistä teknologioista, esimerkiksi data mining, pohjautuvat kypsiin ratkaisuihin, kuten ETL-työkaluihin ja relaatiotietokannanhallintajärjestelmiin (Chen ym., 2012). Analytiikka voidaan jakaa kolmeen eri kategoriaan, niitä ovat kuvaava (*descriptive*), ennustava (*predictive*) ja ohjaava (*prescriptive*). Delen ja Demirkan (2013) kuvaavat näitä kategorioita seuraavalla tavalla:

- Kuvailevaa analytiikkaa kutsutaan useasti liiketoimintaraportoinniksi. Kuvaileva analytiikka tarjoaa tietoa liiketoiminnan suorituskyvyn lisäksi myös sen mahdollisuuksista sekä ongelmista.
- Ennakoivassa analytiikassa dataan käytetään matemaattisia tekniikoita, joilla pyritään löytämään ennakoitavia ja selittäviä malleja. Ennakoivaan analytiikan käytetään erilaisia tiedonlouhinnan (text, web, media) teknologioita ja aikasarjaisia ennusteita, jotka kertovat mitä tulee tapahtumaan ja miksi.
- Ohjaavassa analytiikassa käytetään dataa ja matemaattisia algoritmeja määrittelemään vaihtoehtoisia toimintatapoja tai päätöksiä, niiden tavoitteena on yleensä parantaa liiketoiminnan suorituskykyä. Käytetyt algoritmit voivat pohjautua joko täysin dataan, asiantuntijoiden tietoon tai näiden yhdistelmään, ne tuottavat tietoa siitä, mikä on paras mahdollinen toimintapa tai tarjoavat kattavan valikoiman tietoa päätöksentekijälle.

Big data

Big data määritelmää käytetään kuvaamaan datamääriä, jotka ovat niin suuria ja tyypiltään monimuotoisia, että niiden tallentamiseen ja käsittelyyn vaaditaan perinteisestä datan käsittelystä poikkeavia teknologioita ja toimintatapoja. Vaikka big datan määritelmä vaihtelee lähteestä riippuen, sen tunnistamiseen on olemassa kolme yleisesti tunnistettua ominaisuutta. Nämä ominaisuudet, jotka erottavat big datan tavallisesta datasta ovat; määrä *(volume)*, nopeus (*velocity)* ja monimuotoisuus (*variety)*. Useat eri tutkijat, esimerkiksi Lokhande ja Khare (2015) sekä Sagiroglu ja Sinanc (2013) kuvaavat näitä ominaisuuksia seuraavalla tavalla.

- Määrällisesti big datassa käsitellään dataa, jonka määrän mittaamiseen eivät riitä tera- tai petatavut. Big datassa toimitaan skaalalla, jossa datan analysointiin ja tallennukseen tarvitaan perinteisestä poikkeavia tallennus- ja analysointimetodeja.
- Nopeudella viitataan siihen, kuinka nopeasti uutta dataa syntyy sekä aikaan, missä dataa tulee pystyä analysoimaan. Useissa aikakriittisissä prosesseissa data tulee analysoida vielä sen ollessa liikkeessä, jotta siitä saatu hyöty voidaan maksimoida.
- Monimuotoisuudella viitataan datan eri esiintymismuotoihin. Big data voi olla strukturoitua (esim. relaatiotietokannat), semi-strukturoitua (esim. XML tai JSON) tai strukturoimatonta (esim. sähköpostit, kuvat ja äänitiedostot).

Näiden kolmen perusominaisuuden rinnalle on myös tarjottu myöhemmin erinäistä joukkoa muita ominaisuuksia, kuten *Veracity* (Lokhande & Khare, 2015) ja *Value* (Sharma, S Tim, Wong, Gadia & Sharma, 2014). Näistä ensimmäinen liittyy tiedon luotettavuuteen ja tarkkuuteen, jälkimmäinen siitä saatavaan ar-

voon. Vaikka nämä ominaisuudet ovat big datan kontekstissa tärkeitä, varsinkin terveydenhuollon näkökulmasta, kolme ensimmäistä ominaisuutta riittävät big datan määrittelyyn.

Business intelligence & Analytics

Business intelligence & analytics (BI&A) eli liiketoimintatiedon hallinta ja analytiikka on käsite, joka kattaa laajan valikoiman teknologioita, työkaluja ja prosesseja, joilla dataa kerätään ja joilla siitä tuotetaan liiketoiminnan kannalta hyödyllistä ja analysoitavaa tietoa päätöksenteon tukemiseksi. Informaatioteknologissa käsite liiketoimintatiedon hallinta (BI) vakiintui käyttöön 90-luvulla, ja vuosituhannen vaihteessa liiketoimintatiedon hallinnan yhteyteen lisättiin määritelmä analytiikka. Nykyisin käytetään useimmiten määritelmää business intelligence & analytics eli liiketoimintatiedon hallinta ja analytiikka. Perinteisestä liiketoimintatiedon hallinnasta puhuttaessa käytetään joskus lyhennettä BI&A 1.0 ja liiketoimintatiedon hallinnasta ja analytiikasta lyhennettä BI&A 2.0. Jälkimmäisellä versiolla viitataan siis Web 2.0 teknologioihin ja big data analytiikkaan. Internet of Things, esineiden Internet, johon viitataan useasti nimellä Web 3.0, on myös synnyttänyt käsitteen BI&A 3.0.

Complex event processing

Complex event processing (CEP) järjestelmissä pyritään eri lähteistä tulevien jatkuvien tietovirtojen suodattamisella ja yhdistämisellä tunnistamaan yksittäisiä toistuvia tapahtumia, joiden pohjalta on mahdollista luoda parempi kuva koko järjestelmän toiminnasta. CEP-järjestelmien juuret ovat finanssimarkkinoita varten rakennetuissa järjestelmissä, kuten algoritmeihin pohjautuvassa osakekaupassa. CEP-järjestelmät ovat yleistyneet myös muille toimialoille, joissa voidaan hyötyä reaaliaikaisesta datan käsittelystä ja analysoinnista. Nykyisin CEP-järjestelmiä hyödynnetään esimerkiksi tuotantoteollisuudessa RFID sensorien datan analysoinnissa. Tämän lisäksi CEP-järjestelmiä voidaan käyttää myös tietoverkoissa tapahtuvien tunkeutumisyrityksien tunnistamiseen. Complex event processing edustaa BI&A paradigmaa, jossa dataa ei tallenneta analysointia varten erilliseen tietovarastoon vaan sen analysointi tapahtuu datan ollessa liikkeessä (Chaudhuri ym., 2011; Cugola & Margara, 2012).

Vaikka CEP-järjestelmissä kiinnostuksen painopiste on jatkuvasti liikkuvissa datavirroissa, voidaan niiden analysoinnissa hyödyntää myös relaatio- tai NoSQL tietokantoihin tallennettua dataa. CEP-järjestelmissä kulminoituu myös big dataan yleisesti liittyviä haasteita erityisesti tiedon syntymisen nopeuden ja määrän suhteen, tästä syystä CEP-järjestelmät toteutetaan useasti pilvipalveluna. CEP-järjestelmien toteutuksessa käytetään yleensä yksityisiä pilvipalveluita, koska datavirrat sisältävät yleensä liiketoiminnan kannalta kriittistä ja sensitiivistä tietoa. (Ari, Olmezogullari & Celebi, 2012).

Data warehouse

Tiedon tallentamista erilaisiin relaatiotietokantoihin pidetään yhtenä perinteisen liiketoimintatiedon hallinnan, BI&A 1.0, kulmakivistä (Chaudhuri ym., 2011). Data warehouse, eli tietovarasto, on keskitetty tallennuspaikka datalle, joka yleensä sijaitsee organisaation sisällä useissa eri tietojärjestelmistä tai tiedostoista. Tämä eri muodoissa oleva data muokataan ja tallennetaan relaatiotietokantaan osana ETL-prosessia, jonka jälkeen sitä voidaan hakea SQL kyselyillä. Tietovarasto voi pitää sisällään myös pienempiä osajoukkoja, joita kutsutaan nimellä data mart. Näihin osajoukkoihin on tallennettu yleensä pieni osa koko tietovaraston sisällöstä, joka on relevanttia esimerkiksi vain osalle tietovaraston käyttäjistä.

Internet of Things (IoT)

Internet of Things eli esineiden Internet on termi, johon tänä päivänä törmää lähestulkoon väkisin, jos seuraa IT-alan uutisointia ja keskustelua. Esineiden Internetillä tarkoitetaan nopeasti kasvavaa verkkoon liitettyjen laitteiden, kodinkoneista autoihin, luomaa verkostoa. Web 2.0 teknologiat mahdollistivat uuden datan tuottamisen mm. erilaisten sosiaalisen median palveluiden kautta ja se tapahtui yksittäisten käyttäjien toimesta. Esineiden Internetin aikakaudella uutta dataa syntyy automaattisesti, erilaisten sensorien ja mobiililaitteiden kautta. Esineiden Internetin myötä siirrytään uuteen vaiheeseen myös siinä, miten ja missä iso osa datasta syntyy tulevaisuudessa. (Hu, Wen, Chua & Li, 2014.)

Machine learning

Machine learning, koneoppiminen, on analytiikan osa-alue, jossa analyyttiset mallit luodaan automaattisesti erilaisten tilastollisten ja matemaattisten algoritmien avulla. Koneoppimista voidaan käyttää useissa eri analytiikan sovelluksissa kuten verkkokauppojen hinnoittelussa, hakukoneissa sekä puheen- tai kuvientunnistusta käyttävissä järjestelmissä. Koneoppivia järjestelmiä käytetään myös terveydenhuollossa. Käytössä on mm. järjestelmiä, jotka voivat monitoroida potilastietojen käyttöä ja raportoida mahdollisesti luvattomasta potilastietojen käytöstä organisaation tietoturvavastaaville (Menon, Jiang, Kim, Vaidya & Ohno-Machado, 2014).

Koneoppimisjärjestelmän käyttöönotolle tyypillistä on, että ensin järjestelmää opetetaan käyttämällä harjoitusdatasta koostuvia sarjoja. Tämän manuaalista työtä vaativan opetusvaiheen jälkeen alkaa iteratiivinen prosessi, jossa analytiikkaa suorittava järjestelmä oppii jatkuvasti analysoimastaan datasta. Koneoppimisjärjestelmille ei anneta analysoitavasta datasta selkeää kohdetta vaan sillä pyritään saamaan parempaa käsitystä datan sisältämästä piilotetusta tiedosta.

MapReduce

MapReduce on alun perin Googlen kehittämä ohjelmointimalli verkkosivujen, lokitiedostojen ja palveluiden käyttäjien tuottaman datan analysointiin. MapReduce nimi tulee mallin kahdesta eri funktiosta; Map ja Reduce. Map-funktiota käytetään missä tahansa muodossa olevan datan muuttamiseen avain/arvo pareiksi, jonka jälkeen tästä syntyneet sarjat siirtyvät Reduce-funktion käsiteltäväksi. Reduce-funktio nimensä mukaisesti ”pienentää” saamansa datan esimerkiksi yhdeksi avain/arvo pariksi. MapReduce-malliin pohjautuvat ratkaisut ovat suorituskykyisiä ja sen lisäksi niitä on mahdollista mukauttaa lähestulkoon minkä tahansa strukturoimattoman datan tallentamiseen ja analysointiin. Tämä mukautuvuus on yksi niistä syistä, miksi MapReduce-pohjaiset ratkaisut ovat viime vuosina nousseet suosituiksi. (Chaudhuri ym., 2011)

MapReduce-ohjelmointimalliin pohjautuvista avoimen lähdekoodin ratkaisuista tunnetuin on Apachen Hadoop. Hadoop tarjoaa kustannustehokkaan tavan rakentaa vikasietoisia ja skaalautuvia analyyttisia järjestelmiä. Hadoopin suosiosta kertoo myös se, että suurimmat kaupallisia tietokannanhallintajärjestelmiä toimittavat yritykset (esim. Microsoft, IBM, Oracle) ovat toteuttaneet omat Hadoop pohjaiset ratkaisunsa. (Chen ym., 2012; Lokhande & Khare, 2015; Sharma ym., 2014) Edellä mainittujen ominaisuuksien lisäksi Hadoop mahdollistaa millä tahansa ohjelmointikielellä kirjoitettujen sovellusten ajamisen Mapper tai Reduce tehtävänä, jos ne käyttävät standardeja STDIN ja STDOUT input ja output jonoja. Vaikka MapReduce malliin pohjautuvat ratkaisut sopivat hyvin suurien tietomäärien analysointiin, ne eivät välttämättä sovi I/O-intensiivisten tehtävien suorittamiseen. (Markonis, Schaer, Eggel, Müller & Depeursinge, 2015)

OLAP

OLAP-lyhenne tulee sanoista Online Analytic Processing. OLAP-palveluita käytetään olemassa olevan datan prosessointiin kuutioksi kutsutun, moniulotteisen näkymän luomiseen. Loppukäyttäjä voi suorittaa syntyneisiin moniulotteisiin näkymiin eli kuutioihin analyyttisiä kyselyitä useasta eri liiketoiminnallisesta näkökulmasta. Monet kaupalliset BI-sovellukset hyödyntävät OLAP-teknologioita, sillä ne sopivat erityisen hyvin yritysten myynnin raportointiin ja ennusteiden tekemiseen. Chaudhuri ym. (2011) tiivistävät OLAP-palvelujen keskeisimmät ominaisuudet hyvin; OLAP-palvelut tarjoavat loppukäyttäjälle moniulotteisen (multidimensional) näkymän tietoon ja mahdollistavat sen kokoamisen ja suodattamisen sekä erilaiset pivotointi- ja porautumisoperaatiot.

2.2 Liiketoimintatiedon hallinnan ja analytiikan kehitys

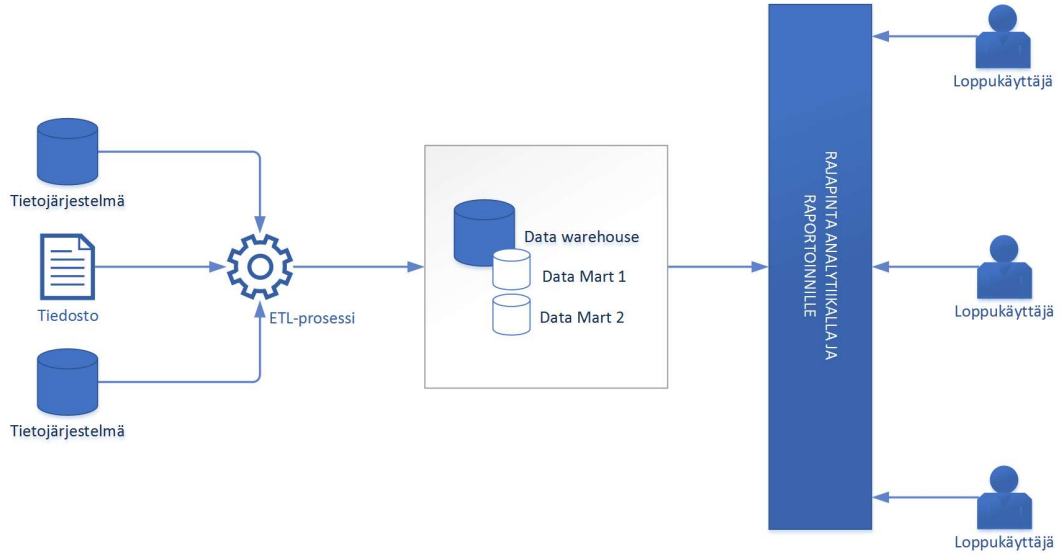
Perinteisen liiketoimintatiedon hallinnan (BI&A 1.0), konvertoinnin ja integraation kulmakiven muodostavat tietovarastot, tietovarastojen osajoukot sekä ETL-prosessit. Niiden juuret löytyvät 90-luvulta ja nämä teknologiat ja prosessit ovat hyvin kypsässä kehityksen vaiheessa. Tästä huolimatta tai kenties tämän ansiosta, näkökulmasta riippuen, BI&A 1.0 kattaa yhä valtaosan yritysten tänä päivänä käyttämästä liiketoimintatiedon hallinnasta ja analytiikasta. BI&A 1.0 arkkitehtuuriin pohjautuvissa järjestelmissä tietoa kerätään lähestulkoon aina useista erillisistä ja vanhoista (legacy) järjestelmistä ja tallennetaan osana ETL-prosessia strukturoidussa muodossa relaationaalisiin tietokannanhallintajärjestelmiin, nämä muodostavat tietovarastoille taustasovellusten (back-end) osuuden. (Chaudhuri ym., 2011; Chen ym., 2012.)

Liiketoimintatiedon hallintaan ja analytiikkaan liittyy myös oleellisesti se, kuinka analysoitava data syntyy. Hu ym. (2014) ovat tunnistaneet tässä kolme eri vaihetta, jotka rinnastuvat liiketoimintatiedon hallinnan ja analytiikan sekä web-teknologioiden kehitykseen. Vaiheessa 1 suurin osa syntyneestä datasta oli tallennettu tietokantoihin ja se oli myös usein strukturoidussa muodossa. Vaiheessa 2 suurin osa datasta syntyi Web 2.0 sosiaalisissa palveluissa. Viimeisessä vaiheessa eli vaiheessa 3 uutta dataa alkaa syntymään automaattisesti ja yhä kiihtyvällä tahdilla esineiden Internetin yleistyessä.

Seuraavassa kuvassa on esimerkki perinteisestä BI&A 1.0 ympäristön arkkitehtuurista. ETL-prosessilla haetaan tietoa eri tietojärjestelmistä ja tiedostoista ja viedään tietovarastoon. Tietovaraston sisällä on pienempiä osajoukkoja, jotka on tarkoitettu organisaation sisällä oleville yksiköille. Loppukäyttäjän ja tietovaraston välissä sijaitsee rajapinta jossa muodostetaan raportteja, OLAP-kuutioita ja muilla tavoin tuotetaan analyysejä. Loppukäyttäjät katsovat jalostettua analyyttistä tietoa esimerkiksi selaimella HTML-pohjaisina raportteina tai taulukkolaskentaohjelmien ja erilaisten ”kojetaulujen” (dashboard) kautta.

Tietovarastoinnin teknologiat kehittyvät yhdessä relaatiotietokantojen hallintajärjestelmien ja palvelinkomponenttien kanssa. Tänä päivänä suurimmissa kaupallisissa tietokannanhallintajärjestelmissä, esimerkiksi Oracle[1] ja Microsoft SQL Server[2], on panostettu muistinvaraisten tietokantateknologioiden kehittämiseen BI- ja OLTP-käytössä. Palvelinkomponenttien, erityisesti keskusmuistin, hintojen laskun seurauksena nykyisin tietokantapalvelimina käytettävissä tietokoneissa voi olla useita teratavuja keskusmuistia. Tämä on tehnyt muistinvaraisista teknologioista houkuttelevan vaihtoehdon analytiikkaan ja muihin äärimmäisen nopeaa suorituskykyä vaativiin tehtäviin.

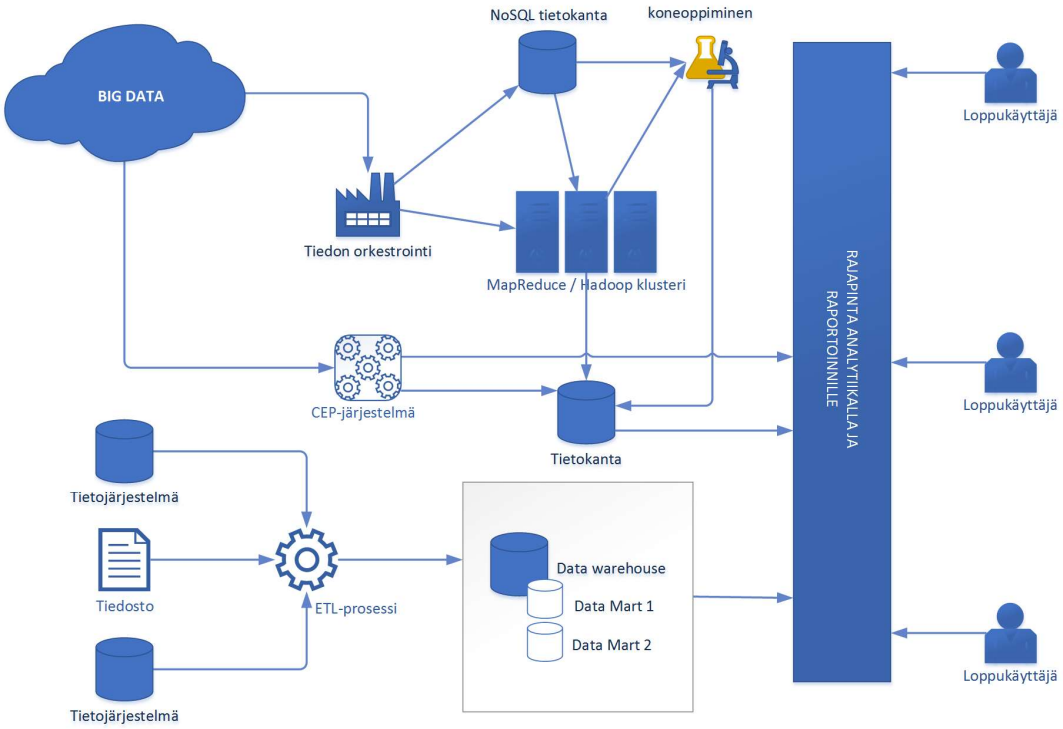
[1] Oracle Database In-Memory (Oracle, 2016)
[2] In-Memory OLTP (Microsoft, 2015)



KUVIO 1 Esimerkki perinteisestä BI&A 1.0 arkkitehtuurista

Monilla toimialoilla, myös terveydenhuollossa, työ on nykyisin yhä useammin liikkuvaa paikkasidonnaisuuden sijaan. Tämän kehityksen myötä useat BI&A-sovellukset tarjoavat mahdollisuutta mobiililaitteiden käyttöön päätelaitteina. BI&A-sovellusten kykyä tarjota näkymää liiketoimintatietoon mobiililaitteilla kutsutaan mobile BI:ksi. Mobile BI:tä pidetään yhtenä tärkeimmistä BI&A-markkinoita mullistavista teknologioista sekä tuotekehityksen ykköskohteista. (Chaudhuri ym., 2011; Chen ym., 2012.) Yksi tärkeimpiä mobile BI-kehityksen mahdollistavia teknologioita on HTML5. Useat BI&A-toimittajat tukevat jo HTML5 versiota omissa tuotteissaan, ja HTML5 tuen yleistyminen eri laitteilla ja selaimilla tulee varmasti vaikuttaman merkittävästi mobile BI-ratkaisujen kehitykseen.

Analytiikan kehityspolku on selvästi big data -painotteinen ja siihen vaikuttavat voimakkaasti myös muut IT-alan trendit, kuten pilvipalvelujen ja palvelukeskeisten arkkitehtuurien kasvava suosio. Uskon kuitenkin, että perinteiset BI&A 1.0 -arkkitehtuurin pohjautuvat ratkaisut tuskin ovat katoamassa vielä lähivuosina vaan ne säilyvät käytössä uusien teknologioiden rinnalla. Seuraavassa kuvassa on esimerkki arkkitehtuurista, jossa uudet analytiikan teknologiat on otettu käyttöön vanhojen BI&A-teknologioiden rinnalle. Kuvaa on selvyyden vuoksi myös yksinkertaistettu käyttämällä vain tässä tutkielmassa esitettyjä teknologioita ja komponentteja. Kuvassa perinteisten BI&A teknologioiden yläpuolella olevat BI&A 2.0 -teknologiat voivat tarvittaessa käyttää hyödyksi myös perinteisiin tietovarastoihin tallennettua tietoa.



KUVIO 2 Esimerkki kehittyneestä BI&A 2.0 arkkitehtuurista

Ne yritykset ja organisaatiot, jotka voivat hyödyntää esineiden Internetin tai sosiaalisen median sisällön analysoinnin tuottamaa dataa, tulevat toimimaan kehityksen keihäänkärkenä. Vaikka tässä tutkielmassa lähestytään analytiikkaa nimenomaan teknologioiden kautta, on tärkeää ymmärtää, että kehittyneen analytiikan käyttöönotossa ei ole kyse pelkästään teknologiakyvykkyydestä. Suunnitteluvaiheessa on teknologiaosaamisen lisäksi yhtä tärkeää ymmärtää ja huomioida liiketoiminnan tavoitteita ja sitä, kuinka analytiikan käyttö voi auttaa niiden saavuttamisessa. Tärkeässä asemassa on myös organisaatiossa vallalla oleva uusien teknologioiden käyttöönottoihin liittyvä kulttuuri.

Vaikka perinteiset analytiikan työkalut ja teknologiat ovat toistaiseksi voimissaan, todennäköisin kehityspolku on, että niiden rinnalle nousevat uuden analytiikan ratkaisut. Kun ne saavuttavat tarpeeksi korkean kypsyysasteen, ne mahdollistavat vanhojen arkkitehtuurien ajamisen alas. Osa näistä teknologioista, esimerkiksi tietokannat, tuskin tulevat poistumaan, vaan ne jatkavat omaa nykyisenkaltaista evoluutiotaan tarjoten uusia ominaisuuksia ja jatkuvasti parantuvaa suorituskykyä analytiikan haasteisiin.

2.3 Liiketoimintatiedon hallinnan ja analytiikan tulevaisuus

Kuten edellisessä luvussa todettiin, analytiikan tulevaisuuteen liittyy keskeisesti big data. Big data -analytiikan ympärille on syntynyt kaksi erilaista paradigmaa; tietovirtojen (stream) ja erien (batch) prosessointi (Chaudhuri ym., 2011; Kambatla ym., 2014; Lokhande & Khare, 2015). Edellä mainitut paradigmat eroavat toisistaan niin käytettävien teknologioiden kuin näiden teknologioiden käyttötarkoitusten osalta. Tietovirtojen prosessoinnin paradigman keskeinen ajatus on, että datan arvo on korkeimmillaan tiedon ollessa mahdollisimman tuoretta. Tietovirtojen prosessoinnissa datan analysointi tapahtuu datan ollessa liikkeessä, kun taas eräprosessoinnin paradigmassa data ensin tallennetaan, jonka jälkeen se analysoidaan.

Tietovirtojen prosessointi sopii hyvin niihin ratkaisuihin, joissa tietoa syntyy useina erilaisina virtoina ja joissa tärkeää on tiedon nopea analysointi ja tuloksiin reagointi. Käyttökohteita terveydenhuollossa ovat esimerkiksi erilaiset elintoimintoja seuraavat monitorit. Tietovirtojen prosessoinnin etuna on, että se ei vaadi laskentatehoa tai tallennuskapasiteettia samalla tavalla kuin erä prosessoinnin ratkaisut. Toistaiseksi eräprosessoinnin paradigma on näistä kahdesta suositumpi ja sen ominaisuuksia on otettu käyttöön myös reaaliaikaisuutta vaativissa ratkaisuissa. (Hu ym., 2014.) Alla olevaan taulukkoon on listattuna keskeisimmät paradigmojen väliset erot.

TAULUKKO 1 Erot erä- ja tietovirtaprosessointien paradigmoissa

	Erä prosessointi	**Tietovirta prosessointi**
Analysoitavan datan sijainti.	Tallennettu strukturoitu, semi-strukturoitu tai strukturoimaton data.	Tietovirta/tapahtumat
Data tallennetaan analysointia varten.	Kyllä.	Ei.
Datan analysointikerrat.	Useita kertoja.	Yleensä kerran.
Analysoinnin kesto.	Pitkä	Millisekunneista sekunteihin.

Big data -analytiikan kehityksessä on useita haasteita, joista yksi keskeisimpiä on tiedon tallentamiseen vaadittava kapasiteetti ja nykyisten tietokannanhallintajärjestelmien suorituskyky ja rajoitukset. Eräät tutkijat, kuten Sharma ym. (2014) sekä Belle ym. (2015), nostavat esille tämän ongelman, joka on seurausta yhä kasvavista datamääristä ja siitä, että dataa syntyy yhä moninaisimmissa eri muodoissa. Big data -analytiikan luomien haasteiden vastaamiseen IT-alalle on kehittynyt täysin uusi toimiala, joka kehittää analyyttisia ratkaisuja, jotka pohjautuvat mm. erilaisiin MapReduce- ja NoSQL-teknologioihin (Sagiroglu & Sinanc, 2013; Sharma ym., 2014).

Toinen big data -analytiikan merkittävistä haasteista on niin sanottu piilotetun big datan ongelma. Eräissä tutkimuksissa, esim. Fan ja Bifet (2013), viitattiin ICD:n tekemään selvitykseen jossa todetaan, että suurin osa mahdollisesti

hyödyllisestä datasta yksinkertaisesti katoaa. Tässä selvityksessä oli mainittu, että tällä hetkellä big datasta hyödynnetään noin 3 %, kun big datasta noin 23 % on potentiaalisesti hyödynnettävissä. Kahdeksi yleisimmäksi syyksi hyödyllisen datan katoamiselle mainittiin sen merkkaamattomuus ja se, että data on strukturoimatonta. (Fan & Bifet, 2013).

Big datan analytiikassa käytetään intensiivistä laskentatehoa vaativia, tiedonlouhintaan tarkoitettuja algoritmeja ja se vaatii poikkeuksetta paljon tallennuskapasiteettia. Tietokannanhallintajärjestelmien kehityksen lisäksi viime vuosina markkinoille on tullut useilta eri toimittajilta tietovarastointiin optimoituja laitteistoja. Nämä laitteistot koostuvat palvelin- ja tallennusjärjestelmistä, jotka on optimoitu suurien tietomäärien varastointiin ja analysointiin, ja niiden käyttö- ja tietokantojenhallintajärjestelmät tulevat esiasennettuina ja valmiiksi määriteltyinä. (Chaudhuri ym., 2011.) Näitä laitteistoja toimittavia yrityksiä on useita, mm. Teradata, SAP, Microsoft, Oracle ja IBM. Vaikka nämä laitteistot on alun perin suunniteltuja käytettäväksi perinteisessä tietovarastoinnissa, monet näistä järjestämistä tarjoavat mahdollisuuden Hadoop-integraation ja siten työkaluja big datan analytiikkaan. Osa laitteistoista, kuten SAP HANA[3], hyödyntää myös muistinvaraisia tietokantoja suorituskyvyn takaamiseksi.

Optimoitujen ja usein kalliiden laitteistojen lisäksi skaalautuvaa suorituskykyä ja tallennuskapasiteettia voidaan tarjota myös pilvipalveluista. Ajattelua, jossa analyyttisen palvelun alusta tuotetaan pilvestä, edistää palvelukeskeisten arkkitehtuurien kasvava suosio. Tämän lisäksi yhä useammin yritysten ja organisaatioiden palveluja tuotetaan sekä tietoa tallennetaan pilvipalveluihin. Pilvipalvelut ovatkin pikkuhiljaa muodostumassa yritysten infrastruktuurin uusiksi kulmakiviksi. (Talia, 2013.) Parin viimeisen vuoden aikana markkinoille onkin tullut useita analytiikkaratkaisuja, jotka toteutetaan kokonaan pilvestä Software-as-a-Service -mallilla kuten Microsoftin Cortana Analytics Suite sekä IBM:n Watson. Cortana ja Watson edustavat lähestymistapaa, jossa kehittyneen analytiikan käyttöönoton lisäksi myös analytiikan käyttämisen kynnystä halutaan laskea, esimerkiksi mahdollistamalla analysoitavien kysymysten esittäminen luonnollisilla kielillä (Demirkan & Delen, 2013).

Big data -ilmiö liittyy myös olennaisesti meneillään olevaan avoimen lähdekoodin vallankumoukseen. Useat suuret kansainväliset yritykset kuten Facebook, Twitter ja Microsoft osallistuvat avoimen lähdekoodin projektien kehitykseen. Yksi näistä analytiikan kannalta merkittävistä avoimen lähdekoodin projekteista on R-kieli. R-kieli on suunniteltu erittäin suurien tietomäärien analysointiin ja on tällä hetkellä integroitu mm. Microsoft SQL Server 2016 versioon. Jo aiemmin mainittu MapReduce on toinen hyvä esimerkki analytiikan kannalta keskeisestä avoimen lähdekoodin teknologiasta.

Analytiikan kehittyminen nykyiselle tasolleen on kestänyt lähestulkoon kaksi vuosikymmentä. Kuluvan vuosikymmenen aikana olemme kuitenkin todistaneet useiden uusien ja mullistavien teknologioiden syntyä ja kehitystä, joten on luonnollista olettaa, että myös analytiikan kehitys tapahtuu samanlaisessa, alati kiihtyvässä tahdissa. Chen ym. (2012) toteavatkin, että vaikka olemme

[3] SAP HANA (SAP, 2016)

vasta todistamassa BI&A 2.0 teknologioiden kypsymistä, olemme jo BI&A 3.0 teknologioiden osalta ovenkynnyksellä ilman täyttä varmuutta siitä, mitä uusia ja mullistavia teknologioita tämä kehitys tuo mukanaan.

3 LIIKETOIMINTATIEDON HALLINTA JA ANALYTIIKKA TERVEYDENHUOLLOSSA

Tässä luvussa tarkastellaan terveydenhuollon organisaatioiden suunnittelemia ja toteuttamia kehittyneen analytiikan ratkaisuja. Luvussa ensimmäisenä käydään läpi liiketoimintatiedon hallinnan ja analytiikan mahdollisuuksia ja haasteita terveydenhuollossa yleisellä tasolla. Tämän jälkeen perehdytään maailmalla toteutettuihin analytiikan ratkaisuihin ja lopuksi esitetään arvioita siitä, millaisia analytiikan ratkaisuja voidaan toteuttaa Suomessa. Luvussa vastataan siis tutkielman kysymykseen kuinka terveydenhuollon organisaatiot maailmalla hyödyntävät analytiikkaa ja kuinka analytiikkaa voidaan hyödyntää suomalaisessa terveydenhuollossa.

Terveydenhuollon menoista suurin osa syntyy prosentuaalisesti pienen potilasmäärän hoitokustannuksista. Suomessa meneillään olevan APOTTI-hankkeen[4] yhteydessä on julkisuudessa esitetty väite, että noin 10 % terveydenhuollon asiakkaista aiheuttaa 80 % sen kuluista. Yhdysvalloista esitetty vastaava arvio on, että noin 5 % asiakkaista aiheuttaa 50 % terveydenhuollon kuluista (Bates, Saria, Ohno-Machado, Shah & Escobar, 2014). McKinsey Global Institute on kansainvälinen konsultointiyritys, jonka tekemän selvityksen mukaan kehittyneellä analytiikalla voidaan saavuttaa Yhdysvalloissa vuosittain yli 300 miljardin dollarin ja Euroopassa arviolta 140 miljardin lisäarvo terveydenhuollon toimialalla. Tähän selvitykseen viitataan useissa eri lähteissä, mm. Belle ym. (2014) ja Kambatla ym. (2014), kun analytiikan potentiaalia ja hyötyjä tuodaan esille terveydenhuollossa.

Yhdysvalloissa yhtenä suurimpana terveydenhuollon kulujen aiheuttajana mainitaan useissa eri lähteissä potilaiden uudelleenkäynnit. Uudelleenkäynniksi määritellään ne tilanteet, joissa potilas on kotiutettu mutta joutuu palaamaan hoitoon seuraavan 30 päivän aikana. Nämä uudelleenkäynnit ovat erittäin yleisiä ja myös kalliita terveydenhuollon kannalta. Hoitokulujen laskemisen lisäksi uudelleenkäyntien vähentämisellä on todettu olevan positiivisia vaikutuksia myös sairaalloisuuden ja kuolemantapausten määrissä. Useissa tutki-

[4] APOTTI-hanke (APOTTI, 2016)

muksissa, mm. Bates ym. (2014), on esitetty arvio siitä, että jopa yksi kolmasosa näistä uudelleenkäynneistä olisi analytiikan keinoin estettävissä.

Potilastietoa on tallennettu tietojärjestelmiin useiden vuosikymmenien ajan, niin Suomessa kuin muualla maailmassa. Siksi on luonnollista odottaa, että kaikesta tästä kerätystä datasta voidaan löytää vastauksia siihen, kuinka terveydenhuollon toimintaa voidaan parantaa ja tehostaa. Bates ym. (2014) ehdottavat artikkelissaan myös huomattavasti laajempaa eri tietolähteistä saatavan datan integraatiota näiden korkeita kustannuksia aiheuttavien potilaiden tunnistamiseen. Heidän mukaansa dataa voisi kerätä esimerkiksi potilaan mielenterveyteen ja sosioekonomiseen tilaan liittyvistä yksityiskohdista, kuten avioliitosta tai asumisjärjestelyistä. Yhteiset sosiaali- ja terveydenhuollon tietojärjestelmät ovat askel suuntaan, joka mahdollistaisi tämän suuntaista analytiikan kehitystä.

Kuten aikaisemmin on todettu, terveydenhuoltoa pidetään yhtenä niistä toimialoista, joilla on mahdollista saavuttaa suurimpia hyötyjä kehittyneen analytiikan käytöstä. Tämän lisäksi se on myös toimiala, jolla on tällä hetkellä nopeimmin kasvavat datamäärät. Erään arvion mukaan globaalilla tasolla tallennetun potilasdatan määrä oli vuonna 2011 noin 150 eksatavua ja uutta dataa syntyisi vuodessa noin 1,2–1,4 eksatavua. (Kambatla ym., 2014). Terveydenhuollon dataa syntyy siis huomattavia määriä, yli miljoona teratavua vuodessa. Chen ym. (2012) määrittelevät kahdeksi tärkeimmäksi terveydenhuollon datan lähteeksi geeniperimälähtöisen (esim. genotyypitys ja geenieksprессio) ja asiakaslähtöisen (esim. potilas-, vakuutus- ja lääkitystiedot) datan.

Erityisesti big dataan kohdistuvan analytiikan potentiaalia terveyden ja hyvinvoinnin toimialalla korostetaan useissa tutkimuksissa ennakoivasta hoidosta puhuttaessa. Tämän lisäksi analytiikka nousee esille, kun terveydenhuollon asiakasta pyritään ohjaamaan kohti elämäntyyliä ja käyttäytymismalleja, jotka tukevat hyvää terveyttä (Belle ym., 2015; Kambatla ym., 2014). Ennalta ehkäisevän hoidon aloittaminen on normaalisti haastavaa, sillä terveyden kannalta haitallisten riskien ja näiden yhdistelmien tunnistaminen on ihmiselle hankalaa. Useilla sairauksilla on kuitenkin tunnisteita, joita analysoimalla voidaan minimoida näitä riskejä ennen niiden realisoitumista sairauksiksi. Analytiikalla pyritään myös tukemaan nykyistä potilaskeskeisempää toimintamallia. Potilaskeskeisessä toimintamallissa asiakkaalle tuotetaan henkilökohtaisia, tarpeen mukaan kohdennettuja terveydenhuollon palveluita ja suosituksia, samalla kun asiakas on aktiivisesti mukana oman hoitosuunnitelmansa suunnittelussa. (Chawla & Davis, 2013.)

Potilaskeskeisessä toimintamallissa tärkeässä roolissa on asiakaskohtaisten hoitosuunnitelmien luominen. Tämän onnistumisen yhtenä oleellisena vaatimuksena pidetään analyyttisen sovellusalustan luomista, jonka avulla voidaan potilaan terveyteen liittyvä data koota useista eri lähteistä analysointia ja riskien arviointia varten. Toteutuksessa olisi hyötyä myös muiden potilaiden terveystiedoista johdetun syvällisen tiedon sekä strukturoidun ja strukturoimattoman datan keräämisestä kliinisistä ja ei-kliinisistä modaliteeteistä. Tämän rat-

kaisun avulla voitaisiin luoda nykyistä parempia mahdollisuuksia ymmärtää erilaisia sairauksia ja niiden etenemistä. (Belle ym., 2015; Chawla & Davis, 2013.)

3.1 Terveydenhuollon analyyttiset ratkaisut maailmalla

Seuraavaksi esitellään joitakin sellaisia maailmalla toteutettuja terveydenhuollon ratkaisuja, jotka hyödyntävät kehittynyttä analytiikkaa ja siihen liittyviä teknologioita. Esimerkeissä ei ole suomalaisessa terveydenhuollossa toteutettuja ratkaisuja, sillä tutkielmaan materiaalia etsittäessä näistä ei löytynyt julkaistuja tieteellisiä artikkeleita. Terveydenhuollon tietojärjestelmät ovat sisällöllisesti, teknisesti ja lainsäädännöllisesti monimutkaisia kokonaisuuksia. Tutkielman esimerkit on valittu tarkoituksenmukaisesti hyvinkin erilaisista käyttötapauksista ja näillä valinnoilla on pyritty korostamaan niitä lukuisia eri käyttötarkoituksia, joita kehittyneelle analytiikalla voi olla terveydenhuollossa. Tämän lisäksi on pyritty löytämään esimerkkejä ratkaisuista, joita olisi mahdollista toteuttaa myös suomalaisessa terveydenhuollossa.

3.1.1 Analytiikka päätöksenteon tukena

Ensimmäinen esimerkki on Brasiliasta. Siellä käytetään koneoppimista tuottamaan terveydenhuollon ammattilaisille tietoa, jota voidaan käyttää potilaan hoitoon liittyvän päätöksenteon tukena. Brasilialaisessa terveydenhuollossa käytetään menetelmää, jossa jokaisella potilaalle määritellään seurantatasot (Surveillance Level). Nämä seurantatasot sisältävät mm. erilaisia lääketieteellisiä suosituksia liittyen potilaan tarvitsemaan hoitoon ja opastukseen. Seurantatasojen määrittely tapahtuu arvioimalla potilaalla esiintyviä terveysriskejä sekä vertaamalla potilaan lääketieteellistä hoitohistoriaa terveydentilassa tapahtuneisiin muutoksiin. Jokaisen potilaan osalta seurantataso arvioidaan uudelleen aina potilaskäynnin yhteydessä ja tämä vaatii siten terveydenhuollon ammattilaisilta paljon manuaalista työtä sekä tarkkuutta. Automaattista seurantatasojen määrittelyä varten tutkijat kehittivät ja myöhemmin toteuttivat Vila Loboton lääketieteellisessä keskuksessa koneoppimiseen pohjautuvan ratkaisun. (Pollettini ym., 2012).

Koneoppimisen käyttämä data kerättiin käytössä olevasta sähköisestä potilastietojärjestelmästä, johon oli tallennettu tieto potilaalle tehdyistä tutkimuksista, diagnostiset tiedot, tehdyt lääketieteelliset toimenpiteet, henkilökohtaisia tietoja sekä paljon muuta hoidon kannalta oleellista tietoa. Analyyttisen ratkaisun tarjoamassa visualisoinnissa käytettiin hyväksi Google Maps -palvelua, jonka avulla pystyttiin esimerkiksi kohdentamaan kartalle ne alueet, joilla esiintyi tietty seurantataso. Tämä visualisointi mahdollisti mm. erilaisten epidemioiden puhkeamisen ja niiden etenemisen seuraamisen. (Pollettini ym., 2012)

Toinen esimerkki potilaskeskeisen hoidon kehittämisestä ja koneoppimisen hyödyntämisestä on *CARE*-järjestelmä (Collaborative Assessment and Recommendation Engine), jota kehitettiin tuottamaan yksilöityjä sairauksien ris-

kiarvioita. *CARE*-järjestelmä vertaili keskenään samankaltaisten potilaiden terveystietoja ja loi näiden tietojen pohjalta ennusteen todennäköisimmistä sairauksista. Näitä ennusteita voidaan käyttää suunnitellessa nykyistä paremmin terveyttä ylläpitäviä ja ennaltaehkäiseviä hoitostrategioita, esimerkiksi parantamalla terveydenhuollon ammattilaisen ja potilaan välistä dialogia. (Chawla & Davis, 2013).

CARE-järjestelmää koekäytettiin neljän vuoden ajalta kerättyihin, historiallisiin potilastietoihin. Järjestelmässä oli 13 miljoonan potilaan tiedot joihin liittyi yhteensä 32 miljoonaa hoitotapahtumaa. Näistä potilaista luotiin joukkoja, joilla oli yhteinen perussairaus. Sen jälkeen niihin käytettiin yleensä suositusjärjestelmissä käytettyä collaborative filtering metodia luomaan arvioita siitä millaisia sairauksia potilaille todennäköisesti kehittyisi. Chawla ja Davis (2013) totesivat tutkiessaan järjestelmän tuottamia tuloksia, että se pystyi ennakoimaan puhkeavat sairaudet oikein 51 %:lle potilaista. Tutkijoiden mukaan järjestelmän tarkkuus olisi vielä parempi, jos sen käytössä olisi enemmän dataa, kuten perhehistoria, geneettinen data, laboratoriotulokset sekä enemmän tietoa oireista. Saavutettua tulosta voidaan pitää erittäin hyvänä, kun huomioidaan, että käytössä oli vain osa kaikesta siitä tiedosta mitä terveydenhuollon tietojärjestelmissä on.

3.1.2 Analytiikka yksityisyyden suojan parantamisessa

Suomessa on viime vuosina tullut esille useita tapauksia joissa potilaiden tietoja on katsottu luvatta. Esimerkiksi kirjoittamalla Googleen hakusanoiksi ”katseli potilastietoja luvatta”, saa yli 300 hakutulosta. Näistä törkeimmässä tapauksessa kesätyöntekijä oli vuonna 2015 katsellut luvatta yli 500 potilaan hoitotietoja aikansa kuluksi. Näin törkeät yksityisyyden suojan loukkaukset ovat kuitenkin onneksi melko harvinaisia. Potilaan yksityisyyden suojan loukkausten taustalla on monesti uteliaisuus mutta myös tietämättömyys siitä, mikä on sallittua ja mikä ei. Uteliaisuuteen pohjautuvissa tapauksissa potilastiedot kuuluvat usein julkisuudessa esiintyville henkilöille tai potilaille, joiden terveydentilassa on jotain poikkeavaa ja mielenkiintoista. Tietämättömyyteen liittyville tapauksille on yleistä, että potilastiedot kuuluvat perheenjäsenelle tai muulle lähisukulaiselle. (Menon ym., 2014.)

Melkein missä tahansa päin maailmaa potilastietoja koskevat erittäin tarkat tietosuojamääräykset ja niiden katselusta ja muusta käytöstä syntyy auditoinnin mahdollistavia lokitustietoja. Suurin osa potilastietojärjestelmistä kuitenkin vain tallentaa tätä tietoa ja sen analysointi tapahtuu manuaalisesti, joko pistokokeilla auditointidataan tai potilaan esittämän pyynnön seurauksena, yleensä organisaation tietosuojavastaavan toimesta. Tämä metodi on kuitenkin puhtaasti reaktiivinen ja jos mahdollisia yksityisyyden suojan loukkauksia havaitaan, vahinko on jo tapahtunut. Näillä yksityisyyden suojan loukkauksilla on terveydenhuollon organisaatioille erittäin haitallisia vaikutuksia. Pahimmillaan potilaiden luottamuksen menettäminen voi vaikuttaa negatiivisesti tarjottavan hoidon laatuun, jos potilaana oleva henkilö ei ole uskaltanut kertoa kaikista oireista tai vaivoista. (Menon ym., 2014.)

Koneoppimisen käyttämistä luvattomaan potilastietojen käytön tunnistamiseen on tutkittu laajasti. Aiheen kiinnostavuutta selittää osittain siihen selvästi liittyvä tarve big data analytiikalle. Potilastietoihin kohdistuvia tapahtumia syntyy pienissäkin terveydenhuollon organisaatioissa nopeasti ja paljon. Käytännössä ainoat järkevät tavat tunnistaa luvatonta tietojen käyttöä on luoda siihen automatisoitu järjestelmä, joka opettelee erottamaan luvallisen ja luvattoman käytön. Koneoppiminen on hyvä teknologiavalinta tämän järjestelmän toteuttamiseen sillä potilastietojärjestelmät tuottavat runsaasti auditointidataa, jota voidaan käyttää tämän järjestelmän opettamiseen. Täysin automatisoitua järjestelmää ei vielä käytännössä ole tehty, vaan olemassa olevat ratkaisut vaativat jossain määrin loppukäyttäjän toimenpiteitä, joko järjestelmän ehdottamien mahdollisesti yksityisyyden suojaa loukkaavien tapausten tarkempaan tutkintaan tai koneoppivan järjestelmän koulutukseen. (Boxwala, Kim, Grillo & Ohno-Machado, 2011; Menon ym., 2014.)

Yksityisyyden suojaa parantavien koneoppimiseen pohjautuvien analyyttisten järjestelmien tutkimus- ja kehitystyö on jatkuvaa. Boxwala ym. (2011) mainitsevat artikkelissaan muutamia näistä ratkaisuista, joissa luvattoman käytön tunnistamiseksi haetaan dataa mm. eri järjestelmistä kuten Internetin palveluista, tietokannoista tai organisaation ERP-järjestelmistä. Pelkästään luvattoman käytön tunnistamisen ja siitä raportoimisen lisäksi koneoppimista voidaan hyödyntää myös täysin uudenlaisen ymmärryksen saamiseen potilastietojen luvattomasta käytöstä.

Menon ym. (2014) raportoivat artikkelissaan, että vaikka eräässä tapauksessa luvattoman potilastietojen käytön tunnistaminen automatisoidusti manuaalisen työn sijaan todettiin hyväksi parannukseksi, järjestelmän omistajat olivat huomattavasti innostuneempia uusista oivalluksista, joita analytiikalla voitiin tuottaa. Yksi näistä oivalluksista liittyi tiedon visualisointiin tietoturvasta vastaavan henkilön työn tukemiseksi. Järjestelmän antamista tiedoista voitiin luoda potilastietojärjestelmän käyttäjien ja potilaiden välistä vuorovaikutusta kuvaava kartta, joka paljasti ne käyttäjäryhmät, joilla todennäköisimmin esiintyy luvatonta potilastietojen käyttöä. Tämän ymmärryksen saaminen mahdollistaa siirtymisen reaktiivisesta toiminnasta proaktiiviseen ongelmien ehkäisemiseen, esimerkiksi järjestämällä suurimmille riskiryhmille ylimääräisiä tietosuojakoulutuksia.

3.1.3 Analytiikka lääketieteellisessä kuvantamisessa

Lääketieteellinen kuvantaminen on yksi niistä terveydenhuollon alueista, joissa kehittyneellä analytiikalla ja siihen liittyvillä teknologioilla on useita mahdollisia käyttökohteita. Terveydenhuollossa kuvantamista käytetään laajasti. Muutamia esimerkkejä näistä käyttökohteista ovat mammografia, magneetti- ja röntgenkuvaus sekä ultraäänitutkimukset. Lääketieteelliset kuvat ovat myös usein tärkeässä roolissa potilaan hoitoa suunniteltaessa ja diagnooseja tehtäessä. Analyyttisistä järjestelmistä, jotka kykenevät arvioimaan lääketieteellisiä kuvia, on erityisesti hyötyä silloin, kun tutkittavien kuvien määrä on suuri. Esimerkik-

si kerroskuvauksessa voi yhden tutkimuksen aikana syntyä tuhansia kuvia joiden koko voi olla yhteensä satoja megatavuja. Nämä kuvamäärät eivät vaadi vain laajaa tallennuskapasiteettia vaan myös analyyttisen järjestelmän, joka kykenee nopeaan ja tarkkaan analysointiin. (Belle ym., 2015.) Lääketieteellisen kuvantamisen analytiikkaa on luontevaa lähestyä big data -analytiikan kautta, sillä siitä löytyvät kolme big datan tunnistettua ominaisuutta; nopeus, määrä ja monimuotoisuus.

Belle ym. (2015) viittaavat artikkelissaan Markonis ym. (2015) tutkimukseen MapReducen hyödyntämisestä lääketieteellisen kuvien analysoinnissa ja erityisesti siinä saavutetuista hyödyistä analysointiin käytettävän ajan osalta. Esimerkkejä mainitaan kolme; optimaalisten parametrien etsiminen koneoppivalle järjestelmälle, sisältöön perustuva lääketieteellisten kuvien indeksointi ja kolmiulotteinen aallokemuunnosten analysointi. Markonis ym. (2015) raportoivat MapReduce ratkaisujen suorituskyvystä ja niiden tuomista aikasäästöistä seuraavasti. Optimaalisten parametrien löytämiseen vaadittu aika putosi noin 50 tunnista hieman alle 10 tuntiin. Sisältöön perustuvassa kuvien indeksoinnissa järjestelmä pystyi käsittelemään 100,000 kuvaa tunnissa ja aallokemuunnosten analysointiin tarvittava aika putosi 130 tunnista noin 6 tuntiin. Indeksointiin liittyvässä tapauksessa Markonis ym. (2015) havaitsivat, että I/O-intensiivisissä tehtävissä MapReduce järjestelmän suorituskyky ei ollut erityisen vakuuttava. Aallokemuunnosten analysoinnissa käytettiin hyödyksi Hadoopin streaming ominaisuutta, joka mahdollisti valmiin MatLab® -sovelluksen käyttämisen Mapper- ja Reducer-skripteissä hyvin pienien sovellukseen tehtyjen muutosten jälkeen.

MapReduce/Hadoop-järjestelmiä voidaan hyödyntää analytiikan lisäksi myös muilla tavoilla. Yksi näistä ratkaisuista on *MIFAS* (Medical Image File Accessing System), pilviympäristöön suunniteltu ja rakennettu Hadoop-pohjainen palvelu, jonka avulla lääketieteellisiä kuvia voidaan tallentaa, vaihtaa ja jakaa mm. eri sairaaloiden kesken (Yang, Chen, Chou & Wang, 2010). Kuten aikaisemmin on todettu, lääketieteellisen kuvantamisen tuottama data on monimuotoista ja sen tallentaminen perinteisin keinoin haastavaa suurien datamäärien takia. MapReduce/Hadoop tarjoaa tähän ratkaisun HDFS-tiedostojärjestelmän muodossa. Lisäksi HDFS-tiedostojärjestelmään pystytään tallentamaan suuria määriä dataa. Yang ym. (2010) toteavat, että Hadoopin hajautettu tiedostojärjestelmä tarjoaa kuvien tallentamiseen myös luotettavuutta replikoimalla dataa Hadoop-klusterin noodien kesken.

Tämän lisäksi ratkaisu mahdollistaa tiedon lähettämisen http-protokollaa käyttäen, joten sisältöä voidaan katsoa millä tahansa selaimella tai sovelluksella, joka tukee http-protokollaa. Yang ym. (2010) toteavat myös, että järjestelmässä on yksi heikko kohta, eli niin sanottu *single-point-of-failure*. Huolimatta MapReduce-klusterin noodien kyvystä replikoida kuvat ja niihin liittyvät tiedot keskenään korkean saatavuuden takaamiseksi, HDFS-nimi-noodin vikatilanne johtaa siihen, että koko tiedostojärjestelmä menee offline-tilaan. Edellä mainittua riskiä ja sen vaikutuksia voidaan kuitenkin minimoida rakentamalla kuvia

tallentava järjestelmä pilvipalveluita hyväksi käyttäen, tällöin kyetään varmistamaan palvelimien mahdollisimman korkea saatavuus.

3.1.4 Reaaliaikainen analytiikka

Terveydenhuollossa korostuu tiedon oikeellisuuden lisäksi myös sen oikeaaikaisuuden tarve, sillä terveys ja joskus hengissä selviäminen ovat kiinni tunneista tai pahimmassa tapauksessa vain muutamista minuuteista. Vaikka aikaisemmin tutkielmassa on todettu, että eräajoprosessoinnin paradigmana on analytiikassa suositumpi, juuri terveydenhuollon puolelta löytyy useita hoitotyön kannalta kriittisiä käyttökohteita tietovirtojen prosessoinnille. Näitä kohteita ovat esimerkiksi useat kliiniset monitorit ja sensorit, jotka seuraavat potilaan tärkeitä elintoimintoja ja joita käytetään yleensä juuri silloin, kun potilaan terveydentila on kriittinen. Raghupathin ja Raghupathin (2014) mukaan reaaliaikaisuus analytiikassa on terveydenhuollon keskeisimpiä vaatimuksia ja viiveet, joita esiintyy datan keräämisen ja sen analysoinnin väliltä, on poistettava.

Raghupathi ja Raghupathi (2014) viittaavat artikkelissaan IBM:n tekniseen raporttiin[5], jossa on kuvattu muutamia tuotannossa olevia ratkaisuja reaaliaikaisen analytiikan käytöstä ja sen tuomista hyödyistä. Ensimmäinen esimerkki oli North York General Hospitalista, jossa reaaliaikaista analytiikkaa käytetään hoidon vaikuttavuuden parantamisen lisäksi myös paremman ymmärryksen saamiseksi siitä, kuinka sairaala toimii. Sairaalalle kehitetty analyyttinen järjestelmä keräsi ja prosessoi dataa yli 50 eri lähteestä, jotka sijaitsivat useissa erillisissä sisäisissä järjestelmissä. Järjestelmän todettiin parantavan lääkäreiden ja hallinnollisen henkilöstön käsitystä siitä, kuinka sairaala toimii kliinisestä, taloudellisesta ja hallinnollisesta näkökulmasta.

Toinen esimerkki oli Columbia University Medical Centerin käyttämästä järjestelmästä, joka analysoi monimutkaisia korrelaatioita elintoimintoihin liittyvästä tietovirrasta tietyntyyppisestä aivovauriosta kärsivillä potilailla. Reaaliaikainen analytiikka mahdollisti sen, että todennäköiset komplikaatioriskit voitiin havaita jopa 48 tuntia aikaisemmin kuin ennen ja lääkärit kykenivät aloittamaan toimenpiteet komplikaatioiden ehkäisemiseen aikaisemmin. Mahdollisuus reagoida näihin riskeihin nopeasti on komplikaatioiden tapauksessa erittäin oleellista. Potilailla, jotka ovat kärsineet aivovaltimonpullistuman repeytymisestä johtuvasta verenkiertohäiriöstä, esiintyy usein vakavia komplikaatioita toipumisen aikana. Monet näistä komplikaatioista ovat hengenvaarallisia kuten viivästynyt iskemia, joka on fataali noin 20 % tapauksista.

Viimeinen esimerkki reaaliaikaisen analytiikan käytöstä oli The Hospital for Sick Children, jonka Project Artemis järjestelmää käytettiin tunnistamaan mahdollisten sairaalasyntyisten infektioiden puhkeamista pikkulapsilla. Järjestelmä analysoi hoitotyössä käytettävien kliinisten monitorointilaitteiden, jotka voivat suorittaa parhaillaan 1000 mittausta sekunnissa, tuottamaa dataa ja pyrki tunnistamaan potilaiden elintoiminnoissa merkkejä, jotka liittyvät potentiaalisiin infektioihin. Järjestelmän raportoitiin pystyvän parhaimmillaan tunnista-

[5] Data-driven healthcare organizations use big data analytics for big gains (IBM, 2013)

maan infektioiden puhkeamisen jopa 24 tuntia aikaisemmin kuin mihin aiemmin käytetyllä metodilla on pystytty. Projektin todettiin raportissa olevan vielä melko uusi, joten täysin sen tuomia hyötyä ei ole IBM:n mukaan voitu vielä arvioida laajasti.

3.2 Analytiikka suomalaisessa terveydenhuollossa

Suomessa sähköiset potilastietojärjestelmät ovat olleet käytössä ja pakollisia kaikissa julkisissa terveydenhuollon organisaatioissa jo useiden vuosien ajan. Tämän ansiosta suomalaisilla terveydenhuollon organisaatioilla on käytössään potilasdataa jopa usean vuosikymmenen ajalta, joten analytiikassa käytettävää luonnonvaraa niiltä ei puutu. Myös terveydenhuollon toimintatavat ovat muutoksen edessä. Suomessa ja maailmalla on käyty vuosia keskustelua terveydenhuollon tarpeesta muuttua organisaatiokeskeisyydestä potilaskeskeisempään suuntaan. Näkemys siitä, että informaatioteknologia on yksi keskeisimmistä tekijöistä, joka tulee viemään tätä hoitotyön transformaatiota eteenpäin, on laajalti hyväksytty.

Vaikka terveydenhuolto ja sen organisaatiot maailmalla toimivat monilla eri tavoilla, eri potilastietojärjestelmiin kerätty tieto on useasti sisällöltään samankaltaista. Useimmiten se on tallennettu relaatiotietokantoihin ja on yleisesti käytettyjen standardien mukaista (esim. CDA, DICOM, FHIR, HL7 jne.). Tämä datan samankaltaisuus tarkoittaa sitä, että edellisissä alaluvuissa esiteltyjä ratkaisuja voidaan toteuttaa myös suomalaisissa terveydenhuollon organisaatioissa. Chawla ja Davis (2013) toteavat, että esimerkiksi koneoppimiseen pohjautuvan analyyttisen järjestelmän tarkkuus parantuu, mitä enemmän tietoa potilaasta on saatavilla. Suomessa kaikki potilastieto tallentuu sähköisiin järjestelmiin. Kehittyneen analytiikan käyttöönoton kannalta terveydenhuollon organisaatiot ovat siis optimaalisessa tilanteessa. Uusia mahdollisuuksia analytiikan käytölle syntyy lainsäädännön kehityksen myötä. Esimerkiksi sosiaali- ja terveydenhuollon organisaatioiden yhteisissä tietojärjestelmissä voidaan analytiikan tarpeisiin yhdistää potilastietojen lisäksi myös sosioekonomista tietoa, kuten Bates ym. (2014) ehdottavat.

Suomessa Kansallisen Terveysarkiston (Kanta) palvelut tarjoavat myös mielenkiintoisia mahdollisuuksia kehittyneen analytiikan osalta. Sähköinen resepti ja Potilastiedon arkisto pitävät tällä hetkellä sisällään lääketieteellistä tietoa, joka kattaa kaikki julkisen terveydenhuollon asiakkaat. Myös yksityisen terveydenhuollon organisaatiot ovat liittymässä Kanta-palvelujen käyttäjiksi sitä mukaa, kun niiden käytössä olevat potilastietojärjestelmät saavuttavat Kanta-valmiuden. THL:n esittämän arvion mukaan vuoden 2016 aikana suuret ja keskisuuret yksityisen terveydenhuollon organisaatiot tulevat liittymään Kanta-palveluihin.[6] Tämän jälkeen Sähköinen resepti ja Potilastiedon arkisto kattavat lähes koko Suomen väestön.

[6] Yksityinen terveydenhuolto aloittanut Potilastiedon arkistoon liittymiset (THL, 2016)

Kansallisen Terveysarkiston lisäksi sosiaalisen median sovellukset ja palvelut ovat potentiaalinen lähde analysoitavalle datalle. Twitterin ja vastaavien sovellusten tuottamien tietovirtojen analytiikalla voidaan nykyisin pyrkiä tunnistamaan esimerkiksi erilaisten epidemioiden puhkeamista ja etenemistä. Raghupathi ja Raghupathi (2014) käyttävät tästä esimerkkinä Haitin 2010 maanjäristystä ja sitä, kuinka Twitterin syötteiden pohjalta koleran etenemistä pystyttiin seuraamaan samalla tarkkuudella kuin virallisissa raporteissa, jopa kaksi viikkoa ennen niiden julkaisemista. Yhdistämällä esimerkiksi sosiaalisen median tuottamaa dataa eri potilastietojärjestelmistä tai Kanta-palveluista saatavaan dataan potilasmääristä ja läpimenoajoista, voitaisiin esimerkiksi ennustaa hoitohenkilöstön tai lääkkeiden lisätarpeita eri terveydenhuollon organisaatioissa. Kansalaisille voitaisiin samassa yhteydessä tuottaa palveluita, jotka esimerkiksi kertoisivat henkilön alueen terveysasemien ruuhkatilanteista ja ohjata käyntejä niille terveysasemille, joissa on vapaata hoitokapasiteettiä.

4 YHTEENVETO JA JATKOTUTKIMUSAIHEET

Tutkielmassa perehdyttiin big datan ja analytiikan teknologioihin ja käsitteisiin sekä siihen, kuinka näitä on hyödynnetty ja voidaan mahdollisesti hyödyntää terveydenhuollon näkökulmasta. Tutkielmassa esitettiin kaksi tutkimuskysymystä, jotka olivat: ”Kuinka analytiikka on kehittynyt ja mihin suuntaan se on kehittymässä” sekä ”Kuinka analytiikkaa on hyödynnetty terveydenhuollon organisaatioissa maailmalla ja kuinka sitä voidaan hyödyntää suomalaisessa terveydenhuollossa”. Tutkielman toteutustapa oli kirjallisuuskatsaus ja sitä varten perehdyttiin useisiin tieteellisiin artikkeleihin, joissa käsiteltiin big dataa, kehittynyttä analytiikka ja joissa terveydenhuolto mainittiin yhtenä niistä toimialoista, joka voi hyötyä eniten järjestelmiinsä kerätyn datan ja big datan hyödyntämisestä.

Tutkielman toisessa luvussa vastattiin kysymykseen analytiikan kehityksestä sekä analytiikkaan liittyvistä tulevaisuuden kehityssuuntauksista. Toisessa luvussa esiteltiin myös analytiikkaan liittyvää terminologiaa ja teknologioita, niissä puitteissa kuin niitä on kandidaatintutkielmassa mahdollista esitellä. Tutkielman kolmannessa luvussa perehdyttiin kehittyneeseen analytiikkaan terveydenhuollon näkökulmasta. Kolmannessa luvussa esiteltiin myös maailmalla tehtyä tutkimusta ja terveydenhuollon organisaatioiden toteuttamia analyyttisia ratkaisuja ja sovelluksia sekä tuloksia, joita näillä ratkaisuilla on saavutettu. Kolmannen luvun lopussa käsiteltiin analytiikkaa ja sen tarjoamia mahdollisuuksia lyhyesti myös suomalaisen terveydenhuollon osalta.

Tutkielman tuloksina analytiikan kehityksen voidaan todeta, että analytiikan kehitys on big data -painotteinen. Perinteisen liiketoimintatiedon hallinnan ja analytiikan kuitenkin uskotaan ainakin toistaiseksi säilyvän uusien analytiikan paradigmojen ja teknologioiden rinnalla. BI&A 2.0 -teknologioiden saavuttaessa korkeamman kypsyyden asteen, osa perinteisistä BI&A 1.0 -teknologioista tulee oletettavasti katoamaan tarpeettomina. Informaatioteknologian alalla edetään nopeaa kehityksen aikakautta. Entistä suorituskykyisemmät ja kapasiteeteiltään suuremmat palvelinkomponentit, kuten myös relaationaalisten tietokannanhallintajärjestelmien kehitys, nopeuttavat myös uusien analyyttisten ratkaisujen syntyä. BI&A -2.0 teknologiat ovat vielä kehitysvai-

heessa, mutta jo nyt on näkyvissä uusi analytiikan aikakausi, jonka kehitys etenee yhdessä Web 3.0 -teknologioiden kanssa. Web 3.0 -aikakausi tuo mukanaan myös muutoksen siihen, missä ja kuinka analysoitava data syntyy. Tulevaisuudessa verkon käyttäjät ja sosiaalisen median palvelut eivät ole suurin uuden datan lähde, kuten Web 2.0 -teknologioiden aikakaudella. Tulevaisuudessa suurin osa datasta syntyy automaattisesti miljardeista laitteista ja sensoreista, jotka muodostavat esineiden Internetin, Internet of Thingsin.

Tutkielmassa todetaan myös, että big datan ja analytiikan keinoin voidaan terveydenhuollon toimintaa parantaa ja kehittää useilla eri tavoilla. Koneoppivat järjestelmät tuottavat tulevaisuudessa hoitoa ja siihen liittyvää päätöksentekoa tukevaa tietoa. Koneoppivia järjestelmiä voidaan käyttää myös parantamaan potilaiden yksityisyyden suojaa. Lääketieteellisen kuvantamisen tuottamia datamääriä, jotka täyttävät big datan tunnusmerkit, voidaan tallentaa ja analysoida esimerkiksi MapReduce-malliin pohjautuvilla ratkaisuilla. Lääketieteellisten sensorien ja monitorien tietovirtoja voidaan analysoida reaaliaikaisesti ja näiden tuottamaa tietoa voidaan käyttää hoitotyössä, jossa nopea reagointi potilaan terveydessä tapahtuviin muutoksiin on kriittisessä asemassa. Tämä voi tuottaa parempia hoitotuloksia. Pilvipalvelut ja palvelukeskeinen ajattelu IT-alan yleistyvinä trendeinä tuovat kehittyneen analytiikan myös pienien ja keskisuurien terveydenhuollon organisaatioiden saataville.

Terveyden ja Hyvinvoinnin Laitoksen (THL) tilastojen mukaan terveydenhuollon menot jatkavat tasaista kasvua vuosi toisensa jälkeen. Kun huomioidaan tämä tosiasia sekä tutkielmassa esitetyt arviot analytiikan hyödyistä, on mielestäni selvää, että suomalaisessa terveydenhuollossa tulee kehittää terveydenhuollon organisaatioiden toimintatapoja sekä käytössä olevia terveydenhuollon tietojärjestelmiä siten, että jo kerättyä potilasdataa kyetään hyödyntämään nykyistä paremmin. Suomessa terveydenhuollon analytiikan potentiaalia kasvattaa kattavasti käytössä olevat sähköiset potilastietojärjestelmät. Analytiikan vaatimasta datasta ei ole puutetta. Analyyttisiä ratkaisuja ja järjestelmiä suunniteltaessa ja toteutettaessa tulee pyrkiä hyödyntämään Web 2.0 ja Web 3.0 ajan teknologioita ja tietolähteitä, kuten sosiaalisen median palveluita ja esineiden Internetiin kytkettyjä terveydenhuollon käyttämiä lääketieteellisiä laitteita ja sensoreita.

Nykyisellä teknologialla ihmisen ”lähdekoodi”, hänen geeniperimänsä, on mahdollista purkaa murto-osalla niistä kustannuksista ja siitä ajasta, mitä siihen vaadittiin vielä noin kymmenen vuotta sitten. Tämä avaa paljon mahdollisuuksia terveydenhuollossa riskien ennaltaehkäisemisen ja proaktiiviseen toiminnan kehittämiseen, kun voidaan ennustaa tarkemmin esimerkiksi mahdolliset alttiudet erilaisille sairauksille. Olemme tuskin kaukana siitä todellisuudesta, jossa jo syntymässä koko perimämme kartoitetaan ja tämän pohjalta meille rakennetaan henkilökohtainen, elämänkaaremme mittainen hyvinvointisuunnitelma. Tässä potilaskeskeisessä mallissa saamme räätälöityä palvelua terveydenhuollon palvelua, joka tukee laadukasta elämänlaatua ja toimintakyvyn ylläpitoa mahdollisimman pitkälle vanhuuteen.

Edellä kuvatusta utopiasta on kuitenkin huolestuttavan lyhyt matka dystopiaan, sillä tässä kehityksessä on nähtävissä myös huolestuttavia piirteitä esimerkiksi yksityisyyden suojan kannalta. Väistämättä esiin nousee ainakin seuraava kysymys; kenellä on oikeus ja pääsy kaikkeen tähän meistä kerättyyn dataan ja siitä johdettuun tietoon. Yksityisyyden suojan haasteisiin kiinnitettiin huomiota myös useissa tutkielmaa varten luetuissa artikkeleissa, esimerkiksi Kambatla ym. (2014). Maailmalla jo nyt vakuutusyhtiöt tarjoavat asiakkailleen edullisempia hintoja, jos nämä suostuvat asentamaan autoonsa sensorit, joilla seurataan heidän ajotapaansa. Suomessa tätä vaihtoehtoa ei ole tarjolla, mutta Ruotsissa vakuutusyhtiö If on kokeillut vastaavaa mallia käyttäen hyväksi älypuhelinten sensoreita. Entä jos tulevaisuudessa vakuutusyhtiöt voisivat vaatia asiakkailtaan heidän henkilökohtaisen lähdekoodinsa luovuttamista räätälöityjä vakuutuksia varten? Vaikka tämä järjestelmä pohjautuisi ainakin alussa vapaaehtoisuuteen, voidaanko olettaa jokaisen ymmärtävän tämän päätöksen vaikutuksia yksityisyyden suojansa kannalta?

Nykyaikana enemmistö meistä on tottunut siihen, että näennäisesti ilmaisten palveluiden hintana on luovuttaa palveluntarjoajille jatkuvasti uutta dataa itsestämme. Tätä dataa voitaisiin varmasti käyttää hyväksi myös terveydenhuollon palveluiden kehittämisessä. Esimerkiksi monitoroimalla potilaan sosiaalisen median palveluiden ja mobiililaitteiden käyttöä voidaan saada tietoa potilaan käyttäytymisestä, mikä esimerkiksi masennuksesta kärsivien potilaiden osalta mahdollistaisi oireiden havaitsemisen ja niihin puuttumisen hyvissä ajoin. Tämän kaltaisiin ratkaisuihin liittyviä potentiaalisia hyötyjä tasapainottaa kuitenkin yksityisyyden suojaan liittyvät haasteet, kuten Bates ym. (2014) toteavat artikkelissaan. Yksityisyyden suojan lisäksi big dataan ja sen analytiikkaan liittyy muitakin riskejä, esimerkiksi analyyttisiä järjestelmiä vastaan tehdyt hyökkäykset. Useissa lähteissä mainitaan koneoppivien järjestelmien oppimisessa käytettävän datan ”myrkyttäminen”, minkä seurauksena koko järjestelmän toimivuus vaarantuu. Terveydenhuollossa, jossa tiedon oikeellisuus on äärimmäisen tärkeää, tämän kaltaisiin riskeihin tulee suhtautua vakavasti.

Big data ja analytiikka ovat aiheita, joiden ympäriltä voidaan löytää useita mahdollisia tutkimusaiheita, erityisesti terveydenhuollon näkökulmasta. Tutkielmaan materiaalia etsittäessä erityisen silmiinpistävää oli terveydenhuollon analytiikkaan liittyvän suomeksi julkaistun tieteellisen tutkimuksen puute. Mielenkiintoisia ja mahdollisesti hyödyllisiä tutkimusaiheita olisivat esimerkiksi suomalaisten terveydenhuollon organisaatioiden käyttämät BI&A teknologiat sekä mahdolliset esimerkkitapaukset, joissa big dataa ja analytiikkaa on käytetty. Tulevaisuuden kannalta olisi oleellista tutkia myös yksityisyyden suojaa ja sen turvaamista analytiikan kehittyessä, kuten myös kansalaisten palveluihin liittyvien suurien tietovarastojen hyödyntämistä.

LÄHTEET

APOTTI. (2016, 03/29/2016). APOTTI-hanke. Haettu 06/27/2016 osoitteesta http://www.apotti.fi/

Ari, I., Olmezogullari, E. & Celebi, O. F. (2012). Data stream analytics and mining in the cloud. *Cloud Computing Technology and Science (CloudCom), 2012 IEEE 4th International Conference On,* (857-862). IEEE.

Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. (2014). Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs (Project Hope), 33*(7), 1123-1131. doi:10.1377/hlthaff.2014.0041 [doi]

Belle, A., Thiagarajan, R., Soroushmehr, S. M. R., Navidi, F., Beard, D. A. & Najarian, K. (2015). Big data analytics in healthcare. *BioMed Research International, 2015*, 1-16. doi:10.1155/2015/370194

Boxwala, A. A., Kim, J., Grillo, J. M. & Ohno-Machado, L. (2011). Using statistical and machine learning to help institutions detect suspicious access to electronic health records. *Journal of the American Medical Informatics Association : JAMIA, 18*(4), 498-505. doi:10.1136/amiajnl-2011-000217 [doi]

Chaudhuri, S., Dayal, U. & Narasayya, V. (2011). An overview of business intelligence technology. *Communications of the ACM, 54*(8), 88-98.

Chawla, N. V. & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. *Journal of General Internal Medicine, 28*(3), 660-665.

Chen, H., Chiang, R. H. & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165-1188.

Cugola, G. & Margara, A. (2012). Processing flows of information: From data stream to complex event processing. *ACM Computing Surveys (CSUR), 44*(3), 15.

Delen, D. & Demirkan, H. (2013). Data, information and analytics as services. *Decision Support Systems, 55*(1), 359-363. doi:http://dx.doi.org/10.1016/j.dss.2012.05.044

Demirkan, H. & Delen, D. (2013). Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud. *Decision Support Systems, 55*(1), 412-421.

Domo. (2016, 04/18/2016). Data never sleeps 3.0. Haettu 06/30/2016 osoitteesta https://www.domo.com/blog/2015/08/data-never-sleeps-3-0/

Fan, W. & Bifet, A. (2013). Mining big data: Current status, and forecast to the future. *ACM sIGKDD Explorations Newsletter, 14*(2), 1-5.

Hu, H., Wen, Y., Chua, T. & Li, X. (2014). Toward scalable systems for big data analytics: A technology tutorial. *Access, IEEE, 2*, 652-687.

IBM. (2013, 02/05/2013). Data-driven healthcare organizations use big data analytics for big gains. Haettu 07/01/2016 osoitteesta http://www-03.ibm.com/industries/ca/en/healthcare/documents/Data_driven_healthcare_organizations_use_big_data_analytics_for_big_gains.pdf

Kambatla, K., Kollias, G., Kumar, V. & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing, 74*(7), 2561-2573.

Lokhande, S. & Khare, N. (2015). An outlook on big data and big data analytics. *International Journal of Computer Applications, 124*(11) doi:http://dx.doi.org/10.5120/ijca2015905658

Markonis, D., Schaer, R., Eggel, I., Müller, H. & Depeursinge, A. (2015). Using MapReduce for large-scale medical image analysis. *arXiv Preprint arXiv:1510.06937,*

Menon, A. K., Jiang, X., Kim, J., Vaidya, J. & Ohno-Machado, L. (2014). Detecting inappropriate access to electronic health records using collaborative filtering. *Machine Learning, 95*(1), 87-101. doi:http://dx.doi.org/10.1007/s10994-013-5376-1

Microsoft. (2015, 09/14/2015). In-memory OLTP. Haettu 06/17/2016 osoitteesta https://msdn.microsoft.com/en-us/library/dn133186.aspx

Oracle. (2016,). Oracle database in-memory. Haettu 06/17/2016 osoitteesta https://www.oracle.com/database/database-in-memory/index.html

Pollettini, J., Panico, S., Daneluzzi, J., Tinós, R., Baranauskas, J. & Macedo, A. (2012). Using machine learning classifiers to assist healthcare-related decisions: Classification of electronic patient records. *Journal of Medical Systems, 36*(6), 3861-3874. doi:10.1007/s10916-012-9859-6

Raghupathi, W. & Raghupathi, V. (2014). Big data analytics in healthcare: Promise and potential. *Health Information Science and Systems, 2*, 3-2501-2-3. eCollection 2014. doi:10.1186/2047-2501-2-3 [doi]

Sagiroglu, S. & Sinanc, D. (2013). Big data: A review. *Collaboration Technologies and Systems (CTS), 2013 International Conference On,* (42-47). IEEE.

SAP. (2016,). Sap hana. Haettu 06/23/2016 osoitteesta https://hana.sap.com/abouthana.html

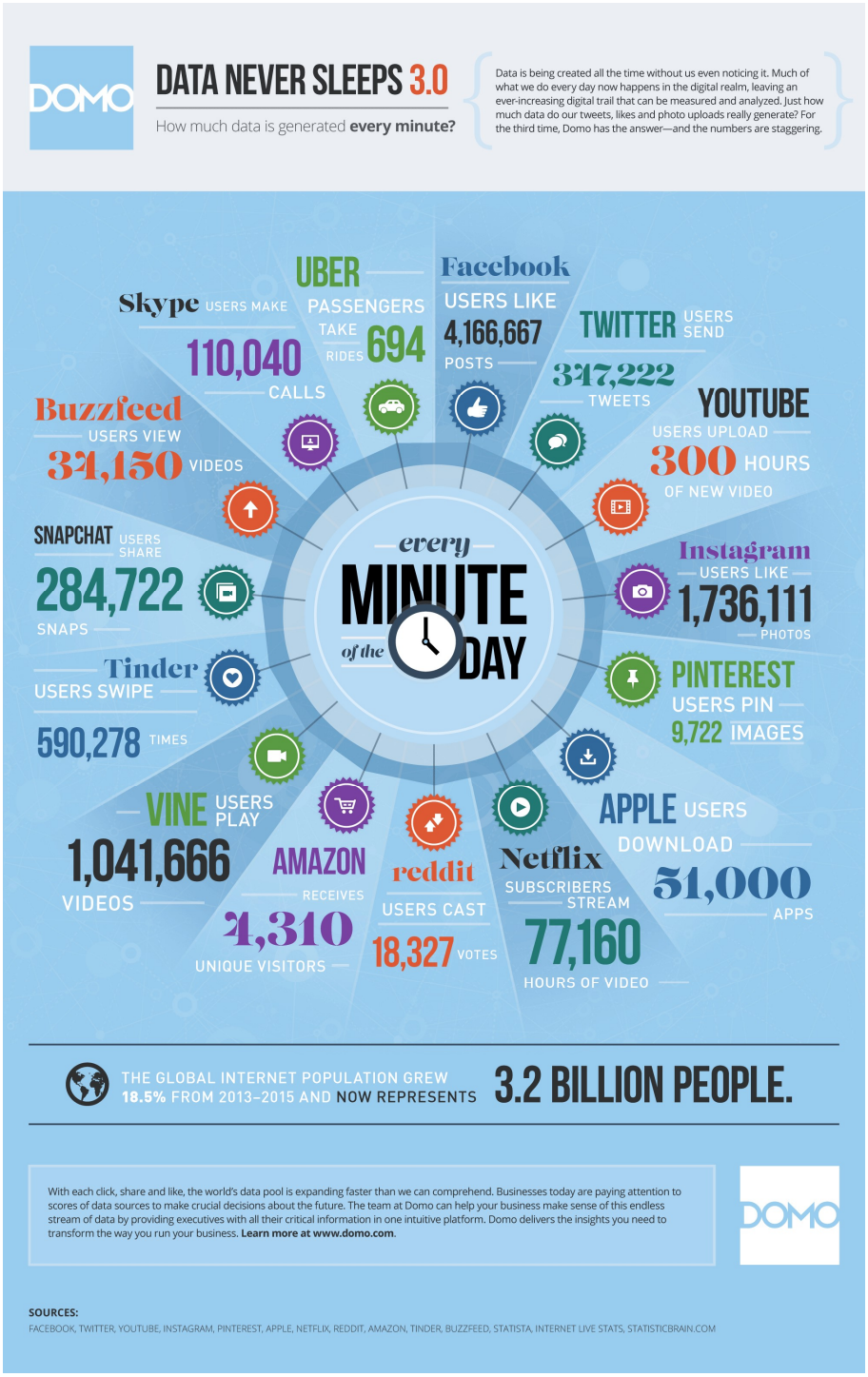
Sharma, S., S Tim, U., Wong, J., Gadia, S. & Sharma, S. (2014). A brief review on leading big data models. *Data Science Journal, 13*, 138.

Talia, D. (2013). Toward cloud-based big-data analytics. *IEEE Computer Science,* , 98-101.

THL. (2016, 02/18/2016). Yksityinen terveydenhuolto aloittanut potilastiedon arkistoon liittymiset. Haettu 07/11/2016 osoitteesta https://www.thl.fi/fi/-/yksityinen-terveydenhuolto-aloittanut-potilastiedon-arkistoon-liittymiset

Yang, C., Chen, L., Chou, W. & Wang, K. (2010). Implementation of a medical image file accessing system on cloud computing. *Computational Science and Engineering (CSE), 2010 IEEE 13th International Conference On,* (321-326). IEEE.

LIITE 1 DATA NEVER SLEEPS 3.0



KUVIO 3 Data never sleeps 3.0, lähde: (Domo, 2016)