Open your mind. LUT.
Lappeenranta University of Technology

TUOTANTOTALOUDEN KOULUTUSOHJELMA

Datan varastointiratkaisun valinta data-analytiikkaa hyödyntävässä yrityksessä

Choosing a platform for data storage in a company utilizing data analytics

Kandidaatintyö

Artturi Jokelainen

Einari Junter

TIIVISTELMÄ

Tekijät: Artturi Jokelainen, Einari Junter

Työn nimi: Datan varastointiratkaisun valinta data-analytiikkaa hyödyntävässä yrityksessä

Vuosi: 2017 **Paikka: Lappeenranta**

Kandidaatintyö. Lappeenrannan teknillinen yliopisto, tuotantotalous.
34 sivua, 6 kuvaa ja 1 taulukko
Tarkastaja(t): Tutkijatohtori, TkT Salla Marttonen-Arola

Hakusanat: analytiikka, tietokanta, pilvipalvelu, NoSQL
Keywords: analytics, database, cloud service, NoSQL

Kandidaatintyö on toteutettu kirjallisuuskatsauksena. Työn tavoite on kartoittaa erilaisia big datan varastointiratkaisuja data-analytiikassa erilaisten yritysten näkökulmasta. Työssä pyritään selvittämään ne päätekijät, joita yrityksen tulee ottaa huomioon valitessaan käytettäviä tietokantaratkaisuja ja niiden hankintamenetelmiä big data -analytiikkaan.

Työssä perehdytään big datan varastointiratkaisuihin niin teknisestä kuin liiketoiminnallisestakin näkökulmasta. Työssä keskitytään etenkin perinteisten relaatiotietokantojen ja nykyaikaisten NoSQL-tietokantojen, sekä ulkoistettujen pilvipalveluna toimivien tietokantapalveluiden ja yrityksen sisäisten IT-infrastruktuuriratkaisujen vertailuun. Huomioon otetaan ratkaisuiden tekninen soveltuvuus big data -analytiikkaan, sekä esimerkiksi kustannukset, tietoturva ja muita liiketoiminnan kannalta oleellisia tekijöitä.

Työn keskeisimpiin tuloksiin kuuluu se, että NoSQL-tietokannat sopivat yleisesti paremmin big data -analytiikkaan. Lisäksi työssä on havaittu, että datan varastoinnin ulkoistaminen pilvipalveluiden avulla voi usein tuoda yritykselle kustannussäästöjä. Työn tuotoksena syntyy taulukosta ja nelikenttämatriisista koostuva graafinen esitys, jonka avulla erilaisten yritysten on mahdollista lähteä valitsemaan tietokantaratkaisua big data -analytiikkaan riippuen yrityksen ja analytiikkaprojektin tavoitteista ja ominaisuuksista.

SISÄLLYSLUETTELO

1 JOHDANTO 3

1.1 Työn tausta, tavoitteet ja tutkimuskysymykset 3

1.2 Työn rakenne ja rajaukset 4

1.3 Työn tutkimusmenetelmät ja aineisto 5

2 BIG DATA JA DATA-ANALYTIIKKA 6

2.1 Big data ja big datan lähteet 6

2.2 Data-analytiikka 8

3 DATAN VARASTOINTI 10

3.1 Datan varastointi data-analytiikassa 10

3.2 Perinteiset relaatiotietokannat ja NoSQL-tietokannat 12

4 PILVIPALVELUT 22

4.1 Pilvipalvelut data-analytiikassa 22

4.2 Pilvipalveluiden tekniset ratkaisut 24

4.3 Pilven palvelu- ja liiketoimintamallit 25

4.4 Pilvipalveluiden valinta 28

5 JOHTOPÄÄTÖKSET 34

LÄHTEET 37

1 JOHDANTO

1.1 Työn tausta, tavoitteet ja tutkimuskysymykset

Työ koskee datan ja analytiikan aiheuttamien kustannusten vertailua erilaisten yritysten tilanteessa datan varastoinnin näkökulmasta. Vertailua tehdään niin erilaisten tietokantatyyppien kuin niiden hankintamenetelmien välillä. Työssä pyritään selvittämään, millaiset erilaiset datan varastointiratkaisut ovat sopivimpia eri yrityksille ja projekteille data-analytiikan kannalta.

Työn aihe on ajankohtainen ja mielenkiintoinen, sillä big data ja siihen liittyvä data-analytiikka ovat saaneet viime vuosina osakseen paljon huomiota ja niiden on uskottu mullistavan monia eri toimialoja. Työn aihe nitoo yhteen tuotantotalouden, tietotekniikan ja liiketoiminta-analytiikan opintoja ja tarjoaa hyvän mahdollisuuden perehtyä tarkemmin siihen, kuinka nämä koulutusalat liittyvät toisiinsa.

Työn tavoitteena on saada selville ne tekijät, joita yrityksen kannattaa huomioida valitessaan sopivinta tietokantatyyppiä data-analytiikan tarpeisiinsa, sekä menetelmää jolla toteuttaa tietokannan hankinta. Työssä pyritään löytämään vastauksia seuraaviin tutkimuskysymyksiin:

1. Miten relaatiotietokannat ja NoSQL-tietokannat soveltuvat big data -analytiikkaan ja mitä eroja niillä on?
2. Mitä erilaisia ratkaisuja yrityksillä on datan varastointiratkaisujen hankintaan data-analytiikassa?

Ensimmäinen tutkimuskysymys käsittelee tietokantoja, niiden tekniikkaa sekä liiketoiminnan kannalta oleellisia eri tietokantaratkaisujen välisiä eroja. Toinen tutkimuskysymys koskee sitä, miten eri yritysten kannattaa toteuttaa datan varastointiratkaisun hankinta. Etsimällä vastauksia näihin kahteen tutkimuskysymykseen voidaan muodostaa kokonaiskuva siitä, miten erilaisten data-analytiikkaa tekevien yritysten kannattaa valita itselleen sopiva ratkaisu datan varastointiin.

1.2 Työn rakenne ja rajaukset

Työ koostuu erilaisia big data -analytiikkaan soveltuvia datan varastointiratkaisuja sekä niihin liittyvää teoriaa käsittelevistä osa-alueista. Työ alkaa johdannolla, jossa määritellään työn rakenne, tutkimuskysymys sekä muut työn pääpiirteet. Johdannon jälkeen kappaleessa 2 on big dataan ja analytiikkaan liittyvää teoriaa käsittelevä osuus. Tässä kappaleessa esitellään big data, sen perusominaisuudet ja yleisimpiä big datan lähteitä. Myös data-analytiikkaa ja sen jaotteluperiaatteita esitellään lyhyesti.

Kappaleessa 3 käsitellään datan varastointiratkaisuja ja niiden soveltuvuutta big data -analytiikkaan. Tässä kappaleessa keskitytään 1. tutkimuskysymyksen käsittelyyn ja huomio pidetään relaatiotietokannoissa ja NoSQL-tietokannoissa sekä niiden välisissä eroissa. Kappaleessa käsitellään myös datan varastointiin liittyvää teoriaa, kuten tietokantoja ja datan rakenteellisuutta. Kappaleen loppupuolella käydään lisäksi läpi, kuinka eri tietokantaratkaisuja sovelletaan reaaliaikaisessa ja ei-reaaliaikaisessa analytiikassa. Kappale 3 päättyy yhteenveto-osuuteen, jossa esitetään taulukon muodossa kappaleessa esiin nousseita 1. tutkimuskysymyksen kannalta oleellisia asioita.

Kappaleessa 4 käsitellään 2. tutkimuskysymystä ja keskitytään erilaisten pilvipalveluratkaisuiden tarkasteluun ja vertailuun datan varastointiratkaisun hankinnan näkökulmasta. Kappaleessa vertaillaan erilaisia pilvipalveluiden teknisiä ratkaisuja ja liiketoimintamalleja sekä niiden eroja. Huomio kiinnittyy niihin seikkoihin, jotka vaikuttavat erilaisten yritysten kohdalla sopivan pilvipalvelumallin valintaan. Lisäksi kappaleessa 4 käydään lyhyesti läpi pilvipalveluihin liittyvää teoriaa sekä niiden kehityksen ja yleistymisen taustalla olleita seikkoja. Kappaleen 4 lopussa on lyhyt yhteenveto-osuus kappaleesta, jossa kiteytetään kappaleessa käsiteltyjä asioita ja esitetään työn tuotoksena nelikenttämatriisi, joka tiivistää 2. tutkimuskysymyksen vastauksia. Työn lopussa, kappaleessa 5, on johtopäätökset. Johtopäätöksissä tiivistetään työn tuloksissa käsitellyt vastaukset tutkimuskysymyksiin ja käydään läpi työn kannalta oleellisia jatkotutkimusaiheita. Johtopäätösten jälkeen tulee lähdeluettelo.

Työ on rajattu siten, että siinä käsitellään tietokantoihin ja pilvipalveluihin liittyviä teknillisiä ja liiketoiminnallisia seikkoja ja käsitteitä. Tietokantojen teknillisten ratkaisujen osalta työssä

keskitytään perinteisten relaatiotietokantojen sekä NoSQL-tietokantojen vertailuun ja muita ratkaisuja sivutaan ainoastaan lyhyesti. Pilvipalveluiden osalta niihin liittyvää tekniikkaa käydään läpi vain pintapuoleisesti, ja huomio kiinnittyy enemmän pilvipalveluiden liiketoimintamalleihin. Työssä ratkaisujen teknillisiä ominaisuuksia käydään läpi sellaisella tarkkuudella, jonka ymmärtäminen ei vaadi aihealueen syvällistä tuntemusta, mutta joka auttaa lukijaa ymmärtämään työn tuloksia.

Liiketoiminnallisen näkökulman osalta työssä keskitytään eri ratkaisujen vertailuun kustannusten kannalta, mutta myös muita liiketoiminnan kannalta oleellisia tekijöitä nostetaan esille. Työssä esimerkiksi sivutaan eri ratkaisuja tietoturvan, henkilöstön koulutuksen sekä liiketoiminnan jatkuvuuden ja riskienhallinnan näkökulmasta.

1.3 Työn tutkimusmenetelmät ja aineisto

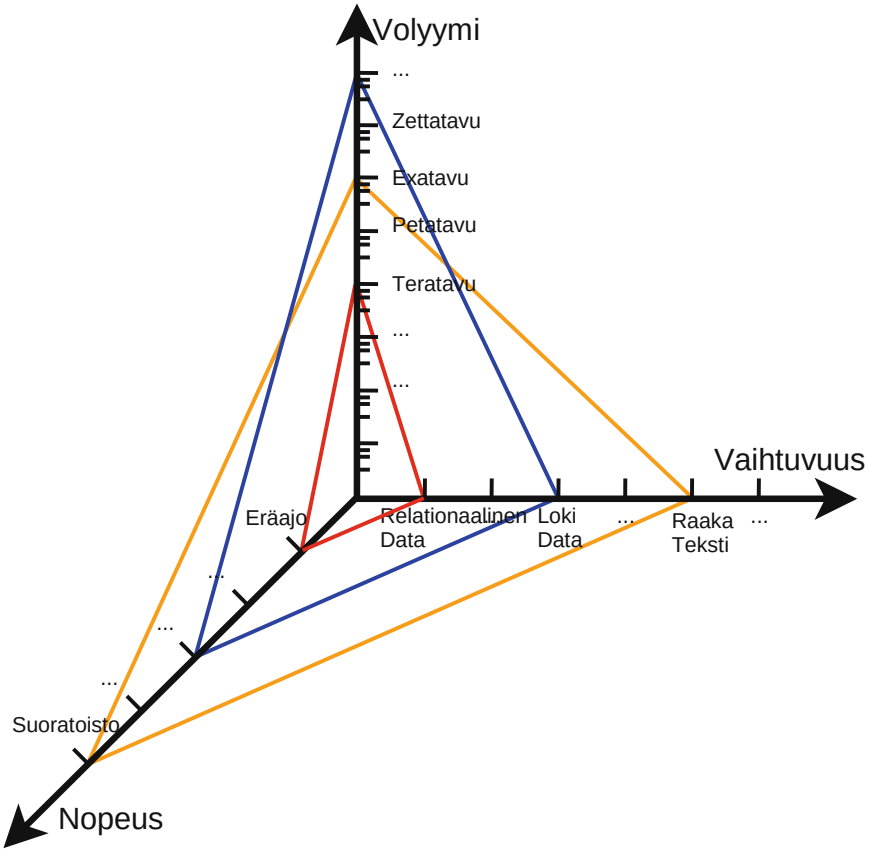
Työ toteutetaan kirjallisuuskatsauksena. Työn aineistona käytetään dataan, data-analytiikkaan sekä pilvipalveluihin ja tietokantoihin liittyviä tieteellisiä kirjoja ja ajankohtaisia artikkeleita. Työn aineiston avulla pyritään luomaan kattava katsaus työn aihepiiriin, eli datan varastointiratkaisuihin ja niiden eroihin big data -analytiikassa. Lisäksi työn aineiston pohjalta luodaan työn tuotos eli vastaus tutkimuskysymyksiin, joka koskee datan varastointiratkaisujen valintaan big data -analytiikassa erilaisten yritysten näkökulmasta.

Työssä keskeisimmät käsitteet pyritään määrittelemään tieteellisen kirjallisuuden avulla, ja ajankohtaisia artikkeleita hyödynnetään alan nykytilanteen ja erilaisten tutkimuskysymyksiin liittyvien teknisten ja liiketoiminnallisten ratkaisujen vertailuun. Työn tutkimuskysymyksiin vastaavien päätelmien tueksi hyödynnetään kuitenkin myös runsaasti tieteellistä kirjallisuutta ajankohtaisten artikkeleiden lisäksi.

2 BIG DATA JA DATA-ANALYTIIKKA

2.1 Big data ja big datan lähteet

Olemassa olevan datan määrä on viime vuosina moninkertaistunut ja kasvanut kiihtyvällä tahdilla (Chen et al. 2014, s. 1). Suuresta ja alati lisääntyvästä määrästä rakenteellisesti erilaista dataa onkin alettu käyttämään käsitettä big data. Big datalla tarkoitetaan dataa, joka ylittää perinteisten tietokantajärjestelmien laskentakapasiteetin. Tämä johtuu esimerkiksi siitä, että dataa on liian paljon, se muuttuu liian nopealla tahdilla tai se on rakenteeltaan yhteensopimatonta perinteisten tietokantojen kanssa. Big dataan liittyy neljä keskeistä ominaisuutta, jotka erottavat sen perinteisestä datasta: suuri volyymi (volume), suuri nopeus (velocity), suuri vaihtuvuus (variety) ja vaihteleva tarkkuus (veracity). (Gupta et al. 2012, s. 42-43)



Kuva 1 Big datan ominaisuudet (Zhao et al. 2014, s. 2)

Yllä olevassa kuvassa 1 on esitetty perinteisen relationaalisen datan sekä big datan eroa kolmiulotteisessa koordinaatistossa, jossa akseleina ovat volyymi, nopeus ja vaihtuvuus. Kuva havainnollistaa, kuinka paljon enemmän tilaa vie esimerkiksi lokitietojen kerääminen (siniset viivat) tai raa'an tekstin jatkuva kerääminen (oranssit viivat) perinteisen relationaalisen datan erä kerrallaan tapahtuvaan keräämiseen verrattuna (punaiset viivat). (Zhao et al. 2014, s. 2)

Big data voidaan jakaa kahteen oleellisesti erilaiseen kategoriaan: lepäävä data (data-in-rest) sekä liikkuva data (data-in-motion). Lepäävä data on dataa, jota on ensin tallennettu ja itse analyysi tapahtuu selvästi erillisellä hetkellä tämän jälkeen. Liikkuva puolestaan kuvaa sellaista tilannetta, jossa dataa analysoidaan heti sen keräämien jälkeen ja sillä pyritään vaikuttamaan tuleviin tapahtumiin reaaliajassa. Olennaiset erot näiden välillä ovatkin juuri analyysin ajankohta verrattuna tallennushetkeen, sekä itse analytiikkaan käytettävät menetelmät. (Gupta et al. 2012, s. 43)

Eräs big datan keskeisimmistä lähteistä on esineiden Internet (Internet of Things, IoT) (Chen et al. 2014, s. 20). Sillä tarkoitetaan keinoja yhdistää esineitä, sensoreita ja muita älykkäitä laitteita ihmisten ja esineiden välisen kommunikaation mahdollistamiseksi. Sen tavoitteena on siis yhdistää virtuaalimaailman informaatio reaalimaailman esineisiin eli integroida tietotekniikkaa esineisiin ja asioihin. Esineiden Internetin tuottamalle datalle on ominaista suuri koko, suuri vaihtuvuus, sitoutuneisuus aikaan ja paikkaan sekä se, että usein vain pieni osa sen avulla tuotetusta datasta tulee olemaan hyödyllistä (Chen et al. 2014, s. 21). (Uckelmann 2011, s. 1-3)

Käsite esineiden Internet syntyi 2000-luvun alussa, kun logistiikka-alalla etsittiin uusia ratkaisuja toimitusten ja tuotteiden seuraamiseen toimitusketjuissa. Myöhemmin käsite on kuitenkin yleistynyt muillakin toimialoilla, ja siitä on tullut yleinen nimitys kehitykselle, jossa esimerkiksi sensoreiden ja automaattisten tunnistamisjärjestelmien avulla liitetään yhä uusia asioita ja esineitä tietoverkkoihin ja tuotetaan niiden avulla dataa myöhempää hyödyntämistä varten. (Uckelmann 2011, s. 2)

Esineiden Internetin yleistymisen taustalla viime vuosina on ollut muun muassa koon ja hintojen pieneneminen RFID-tunnisteissa, kehitys sensoriverkoissa sekä muiden langattomien lait-

teiden ja teknologioiden kehittyminen. Tämänkaltaisen tekniikan kehittymisen johdosta esineiden Internetistä on tullut ajankohtainen asia erilaisille loppukäyttäjille eri toimialoilla. (Uckelmann 2011, s. 2)

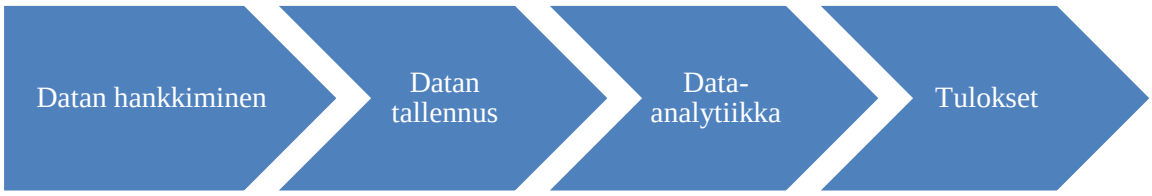
Esineiden Internetin yleistymisen kannalta oleellisimmat haasteet liittyvät tarvittavan infrastruktuurin kehittymiseen sekä tiedon liikkumiseen organisaatiorajojen ulkopuolelle. Monissa olemassa olevissa esineiden Internetin sovelluksissa luotu data pysyy vain organisaation sisällä, ja datan keräämisessä hyödynnettävät tunnistesirut liikkuvat suljetussa kierrossa organisaation omissa toimitusketjuissa. Tulevaisuudessa haasteeksi nouseekin yhä enemmän avoimen kierron järjestelmät, joissa esineet ja data liikkuvat organisaatioiden välisissä verkostoissa. Samaan aikaan tulevaisuudessa tullaan tarvitsemaan avointa, turvallista ja standardisoitua infrastruktuuria, jossa esimerkiksi sensoritieto voi liikkua ja olla haettavissa globaalisti. (Uckelmann 2011, s. 2-3)

Esineiden Internetin lisäksi big dataa syntyy suuria määriä yritysten sisäisestä datasta, kuten asiakastiedoista sekä markkinatiedoista. Myös Internetissä syntyvä data, kuten hakuhistoria ja sosiaalisen median viestit, on merkittävä big datan lähde. Näiden lisäksi big dataa syntyy myös tieteellisen tutkimuksen piirissä ja lääketieteellisissä mittauksissa. (Chen et al. 2014, s. 19-23)

2.2 Data-analytiikka

Data-analytiikka tarkoittaa datan hankintaa, käsittelyä ja muokkaamista datasta saatavan hyödyn maksimoimiseksi. Data-analytiikka voidaan jakaa perinteiseen data-analytiikkaan ja big data -analytiikkaan. Perinteinen data-analytiikka tarkoittaa erilaisten tilastollisten työkalujen käyttöä suuren datamäärän analysointiin (Runkler 2012, s. 2-3). Big data -analytiikka eroaa perinteisestä data-analytiikasta siten, että se keskittyy nimenomaan big datan analysointiin. Big data -analytiikkaan voidaan soveltaa osin samoja työkaluja ja menetelmiä kuin perinteisessäkin data-analytiikassa. Tällöin big datan ominaisuudet, kuten suuri volyymi ja rakenteettomuus, tulee huomioida. Big data -analytiikassa esimerkiksi on otettava huomioon perinteiseen data-analytiikkaan nähden suuremmat vaatimukset laskentateholle ja varastointikapasiteetille. Data-analytiikka on datan arvoketjun viimeinen vaihe yhdessä tulosten kanssa, ja sen tavoite on luoda

käyttökelpoista arvoa datan pohjalta. Datan arvoketjua on hahmoteltu alla olevassa kuvassa 2. (Chen et al. 2014, s. 51-52; Loshin 2013, s. 1-4)



Kuva 2 Datan arvoketju (Casado & Younas 2015, s. 2085)

Big data -analytiikka voidaan jakaa reaaliaikaisuusvaatimuksen perusteella kahteen päätyyppiin: online-analytiikkaan ja offline-analytiikkaan. Online-analytiikalla tarkoitetaan reaaliaikaisesti tapahtuvaa data-analytiikkaa. Sitä hyödynnetään pääsääntöisesti aloilla, joissa analytiikan tuloksia halutaan saada datasta mahdollisimman pienellä viiveellä. Esimerkkejä tällaisista aloista ovat finanssiala ja verkkokauppa. Näillä aloilla luottokorttimaksut ja niihin liittyvät turvallisuustekijät ovat avainasemassa, ja niiden käyttöä on tarkkailtava ja analysoitava jatkuvasti väärinkäyttöyritysten ja huijausyritysten havaitsemiseksi (Gupta et al. 2012, s. 57-58). Online-analytiikassa on käytössä paljon perinteisiin relaatiotietokantoihin perustuvia arkkitehtuuriratkaisuja. (Casado & Younas 2015, s. 2082-2083; Chen et al. 2014, s. 55-56)

Offline-analytiikalla puolestaan viitataan data-analytiikkaan, jonka käyttökohteet eivät vaadi välitöntä vastausaikaa, eli analytiikan ei tarvitse olla reaaliaikaista. Tällaisia käyttökohteita voivat olla esimerkiksi erilaisten suositusalgoritmien tekeminen, koneoppiminen ja tilastollinen tutkimus. Tällöin big data yleensä siirretään lokeista erilliseen ympäristöön, jossa itse analyysi tapahtuu. Kaikkein massiivisimpien datamäärien analyysi kuuluu offline-analytiikkaan. Offline-analytiikassa on käytössä yleisesti mm. Hadoop-pohjaisia arkkitehtuuriratkaisuja, jotka mahdollistavat big datan kustannustehokkaan käsittelyn. Hadoopilla tarkoitetaan läheisesti big dataan liittyvää teknologiaa, joka tarjoaa viitekehyksen erilaisille big datan käsittely- ja varastointiratkaisuille. (Casado & Younas 2015, s. 2082-2083; Chen et al. 2014, s. 16-18, 55-56)

3 DATAN VARASTOINTI

3.1 Datan varastointi data-analytiikassa

Datan keräämisen jälkeen data on varastoitava, ennen kuin sitä voidaan hyödyntää. Datan varastoiminen koostuu kahdesta asiasta: varastointi-infrastruktuurista sekä varastointiin käytettävistä menetelmistä. Tavoitteena on varastoida suuria määriä dataa pitäen järjestelmän toiminta luotettavana ja data aina saatavilla. (Chen et al. 2014, s. 33)

Varastointi-infrastruktuurin kannalta big datan varastoinnissa hyödynnetään paljon hajautettuja varastointijärjestelmiä, ja yksi big datan tuomista haasteista datan varastoinnin kannalta onkin se, kuinka luoda laajoja hajautettuja datan varastointijärjestelmiä big datan varastoinnin ja analysoinnin mahdollistamiseksi. Hajautetuilla varastointijärjestelmillä tarkoitetaan useilla eri palvelimilla sijaitsevia varastointijärjestelmiä, jotka usein sijaitsevat myös fyysisesti useissa eri kohteissa. Hajautetuissa varastointijärjestelmissä keskiöön nousee kolme tekijää: eheys (consistency), saatavuus (availability) ja osioinnin sietokyky (partition tolerance). (Chen et al. 2014, s. 35)

Hajautetun varastointijärjestelmän saatavuudella tarkoitetaan sitä, kuinka hyvin järjestelmä toipuu palvelinten vioittumisesta ja kuinka ison osan ajasta se pystyy palvelemaan asiakkaiden datan luku- ja kirjoitustarpeita. Saatavuutta pyritäänkin parantamaan käyttämällä useita palvelimia samanaikaisesti. Koska hajautetuissa varastointijärjestelmissä hyödynnetään useita eri palvelimia, niihin liittyy se riski, että osa palvelimista hajoaa ja tietoa menetetään. Tämän estämiseksi samasta datasta varastoidaan useita identtisiä kopioita eri palvelimille. Eheydellä tarkoitetaankin hajautetuissa varastointijärjestelmissä sitä, kuinka identtisiä järjestelmän eri palvelimille tallennetut datan kopiot ovat keskenään. Palvelinvirheet ja datan rinnakkainen säilytys voivat vaarantaa datan eheyden hajautetuissa varastointijärjestelmissä. Osioinnin sietokyvyllä puolestaan viitataan hajautetuissa varastointijärjestelmissä siihen, kuinka hyvin järjestelmä kestää eri palvelinten välisiä verkkoyhteysvirheitä. (Chen et al. 2014, s. 35)

Vuonna 2000 esitetyn CAP-teoreeman mukaan hajautettu varastointijärjestelmä voi saavuttaa vain kaksi edellä esitellyistä kolmesta ominaisuudesta. Tämän vuoksi erilaisiin käyttötarkoituksiin on luotu erilaisia hajautettuja varastointijärjestelmiä, joissa jokaisessa on tingitty yhdestä ominaisuudesta kahden muun hyväksi. Esimerkiksi CA-järjestelmät tinkivät osioinnin sietokyvystä taatakseen käyttäjille eheyden ja saatavuuden. Kuitenkin heikon osioinnin sietokyvyn vuoksi niitä ei juuri käytetä usean palvelimen järjestelmissä, vaan ainoastaan pienissä yhden palvelimen järjestelmissä. CP-järjestelmät puolestaan tinkivät saatavuudesta taatakseen osioinnin sietokyvyn ja eheyden. Niitä käytetäänkin kevyen kuormituksen järjestelmissä, joissa datan tarkkuuden on oltava korkea. Esimerkiksi markkinadataa tarjoavissa palveluissa hyödynnetään usein CP-järjestelmiä. AP-järjestelmät taas tinkivät eheydestä saatavuuden ja osioinnin sietokyvyn varmistamiseksi, jolloin järjestelmä kestää suuriakin kuormia mutta datan eheys voidaan taata vasta tietyn viiveen jälkeen. Käytännössä tämä tarkoittaa sitä, että muuttunut tieto päivittyy kaikille palvelimille viiveellä, ja sen ajan tieto voi poiketa riippuen siitä, miltä palvelimelta se luetaan. AP-järjestelmiä esiintyy muun muassa internetin sosiaalisen median palveluiden taustalla. (Chen et al. 2014, s. 35-37; Zhao et al. 2014, s. 23)

Datan varastointimenetelmät rakennetaan edellä kuvatun varastointi-infrastruktuurin päälle. Datan varastointimenetelmänä käytetään yleisesti tietokantoja. Tietokanta on tietokoneistettu järjestelmä, jonka avulla informaatiota säilytetään ja jonka avulla informaatio on helposti saatavilla silloin, kun sitä tarvitaan. Tietokantojen avulla voidaan säilyttää suuria määriä informaatiota halutussa muodossa. Tietokanta voi olla esimerkiksi yksittäisen ohjelman tai käyttöjärjestelmän datan tallennusväline, tai se voi palvella laajempaa kokonaisuutta, kuten yrityksen asiakastietojärjestelmää (CRM). (Chen et al. 2014, s. 33, 37-38; Foster & Godbole 2014, s. 3-8)

Datan rakenteellisuus

Datan rakenteellisuudella tarkoitetaan sitä, että datan sisältämän informaation muoto voidaan määritellä etukäteen. Tietotekniikassa tämä tarkoittaa tiedon pilkkomista esimerkiksi riveihin, taulukoihin tai arvopareihin, jotka noudattavat aina samaa määriteltyä muotoa. Mikäli datan sisällölle ei voida määritellä etukäteen rakennetta, jota se yksiselitteisesti noudattaa on kyseessä rakenteetonta dataa. Yleisesti tietojärjestelmät ovat historian aikana enemmän tuottaneet rakenteellista dataa, sillä ne ovat liittyneet koneisiin tai liiketoimintoihin, jotka vaativat tarkkuutta (esimerkiksi finanssiala ja pankkidata) ja täten on määriteltävä etukäteen. Erityisesti internetin

ja web-sisällön yleistyminen on lisännyt saatavilla olevaa rakenteetonta dataa, josta yleisintä ovat esimerkiksi sähköpostiviestit, kuvat, videot ja sosiaalisen median sisältö (Moniruzzaman & Hossain 2013, s. 2). Näille on hankalaa tai mahdotonta määritellä etukäteen tiettyä rakennetta, sillä niidet sisältö on usein ihmisen tuottamaa ja yksittäisten datan lähteiden välillä vaihtelu on suurta. Jopa 80 prosenttia yrityksissä olemassa olevasta tiedosta on tutkimusten (Grimes 2008) mukaan rakenteetonta. Datan rakenteellisuudella on suuri merkitys valittaessa sopivaa tietokantaa sen varastointiin, sillä ratkaisut on optimoitu toimimaan paremmin joko rakenteellisen tai rakenteettoman tiedon kanssa. (Beach 2014, s. 47; Chen et al. 2014, s. 63)

3.2 Perinteiset relaatiotietokannat ja NoSQL-tietokannat

Perinteiset tietokannat on rakennettu SQL-kyselykielen ympärille (Structured Query Language), joka pohjautuu relaatiomalliin. Näille tietokannoille on ominaista ennalta tarkasti määritelty tiedon rakenne ja sen väliset suhteet. Erityisesti perinteiset SQL -tietokannat sopivat projekteihin, joissa vaaditaan datan ja analyysien yhtenäisyyttä ja eksaktius on olennaista. Relaatiotietokannat ovat olleet pitkän aikaa huomattavasti muita tietokantoja suositumpia, sillä niiden vahvuudet ovat sopineet yrityksissä olevan datan analysoinnin tarpeisiin. (Gupta et al. 2012, s. 42-47)

Viime vuosina on käynyt selväksi, että perinteiset relaatiotietokannat eivät kykene selviytymään kaikista haasteista, joita big datan yleistyminen on tuonut. Niiden puutteisiin big datan varastoinnissa lukeutuu joustamattomuus, yhteensopimattomuus rakenteettoman tiedon kanssa sekä liian heikko tuki massiiviselle määrälle dataa (Casado & Younas 2015, s. 1). Tästä syystä niiden rinnalle ja tilalle on ilmestynyt uusia, muuhun kuin relaatiomalliin perustuvia tietokantaratkaisuja, kuten erilaiset NoSQL-tietokannat. (Chen et al. 2014, s. 38; Gupta et al. 2012, s. 42-47)

Yrityksen tarpeisiin sopivan tietokannan valinta perustuu siihen, millaisia ominaisuuksia käsiteltävällä tiedolla on. Datan keräämisen yleistyessä tilanne on usein se, että sekä rakenteellista että rakenteetonta tietoa on käsiteltävänä paljon. Tietokannan valinnassa vaikuttavia tekijöitä ovat usein myös olemassa oleva yrityskulttuuri ja osaaminen tietokantoihin liittyen, sekä kuinka syvällisesti ohjelmistot on mukautettu toimimaan nykyisillä tietokantaratkaisuilla. Useimmiten

yksittäinen ratkaisu ei riitä täyttämään kaikkia data-analysoinnin tarpeita, vaan usean luonteeltaan erilaisen datan analysointiin vaaditaan erilaiset työkalut. Analyysin näkökulmasta tämä tarkoittaa vaihtelua reaaliaikaisuuden ja myöhemmän analyysin välillä, sekä käsiteltävän datamäärän suuruutta ja hajanaisuutta. Millisekuntien vasteaikaa vaativat reaaliaikaisiin muutoksiin perustuvat analyysit vaativat tietoteknisesti hyvin erilaista lähestymistä, kuin analyysi joka käy läpi valtavan määrän aikaisemmin tallennettua tietoa. (Levin 2015)

Perinteisten relaatiotietokantojen vahvuudet ja heikkoudet

Perinteisten relaatiotietokantojen etuihin kuuluu ennen kaikkea se, että ne tarjoavat erittäin houkuttelevan käyttöliittymän datan hallintaan ja hakemiseen. Niissä on hyvin yksinkertaiset ja helposti käytettävät työkalut datan hakemiseen ja tarkasteluun. Lisäksi ne ovat vuosikymmenien kehityksen ansiosta hyvin vakaita ja tarjoavat erittäin monipuolisia ominaisuuksia. (Zhao et al. 2014, s. 42-44)

Relaatiotietokantojen suuri vahvuus on vahva tuki transaktioille, eli siirtymisille yhdestä tilasta toiseen ilman, että mitään haluttua toimintoa jää toteuttamatta. Transaktio voi olla esimerkiksi tiedon päivittämistä, lisäämistä ja poistamista, tai mikä tahansa muu toiminto joka muuttaa sen sisältöä. Tämä tuki takaa sen, että tietokanta on aina eheässä ja halutussa tilassa, mikä on erittäin tärkeää sen luotettavuuden kannalta erityisesti yrityksen liiketoiminnan kannalta kriittisissä järjestelmissä. (Zhao et al. 2014, s. 44)

Koska relaatiotietokannat ovat tekniikkana hyvin vanha, on kaikilla toimialoilla saatavilla useita toimittajia ja asiantuntijoita relaatiotietokannan hankintaan ja ylläpitoon, mikä on tietokantaa hankkivan yrityksen kannalta erittäin oleellinen tekijä. Relaatiotietokantoihin on helppo saada käyttötukea toimittajilta, ja yrityksen tarvitsemia asiantuntijoita on saatavilla työmarkkinoilla runsaasti toimialasta riippumatta. (Zhao et al. 2014, s. 44)

Relaatiotietokantojen keskeinen heikkous etenkin big dataa käsittelevän yrityksen näkökulmasta on niiden heikko skaalautuvuus (Casado & Younas 2015, s. 1). Kun järjestelmän skaalautuvuuden rajat on saavutettu, on niiden laajentaminen hyvin hidasta ja kallista. Tämä onkin yleinen syy siihen, miksi moni yritys tänä päivänä etsii ja kiinnittää paljon huomiota muihin tietokantaratkaisuihin (Moniruzzaman & Hossain 2013, s. 10-11). (Zhao et al. 2014, s. 42)

Relaatiotietokantajärjestelmien säätäminen sekä ylläpito on hankalaa ja kallista. Tämä johtuu siitä, että relaatiotietokantojen ylläpito vaatii paljon osaamista ja kalliiden ammattilaisten työpanosta. Relaatiotietokannan ylläpitokustannukset voivat muodostaa huomattavan osuuden tietokannan kokonaiskustannuksista, eikä relaatiotietokannasta saada ulos hyvää suorituskykyä ilman koulutettuja ammattilaisia. (Zhao et al. 2014, s. 43)

NoSQL-tietokannat

NoSQL-tietokannoilla (Not Only SQL) tarkoitetaan kaikkia tietokantoja, jotka eivät perustu perinteiseen relaatiomalliin. Poiketen relaatiotietokannoista ne eivät vaadi etukäteen määriteltyä skeemaa datan rakenteelle ja täten ne skaalautuvat paremmin horisontaalisesti. Horisontaalisella skaalautuvuudella tarkoitetaan tietokannan kapasiteetin kasvattamista uusien rinnakkaisten palvelimien lisäämisen avulla sen sijaan, että kehitettäisiin olemassa olevia palvelimia tehokkaammiksi. (Chen et al. 2014, s. 38; Zhao et al. 2014, s. 23)

NoSQL-tietokannat sopivat tapauksiin, jolloin tietokantaan on tarvetta tehdä paljon muutoksia nopealla tahdilla. Big datan sisältämän rakenteettoman tiedon määrän kasvu on kasvattanut kiinnostusta NoSQL-tietokantoja kohtaan. Tietokannat itsessään ovat kehittyneet Internetin suurimpien toimijoiden, kuten Googlen, Amazonin ja Facebookin ajauduttua tilanteeseen, jossa perinteiset relaatiotietokannat eivät pystyneet enää suoriutumaan käyttäjien, järjestelmien ja sensoreiden luomasta massiivisesta datan määrästä pilvipalveluissa. (Chen et al. 2014, s. 38; Moniruzzaman & Hossain 2013, s. 1)

NoSQL-tietokantojen yleistymisen taustalla on vaikuttanut perinteisten relaatiotietokantojen riittämättömyys big datan käsittelytarpeisiin. NoSQL-tietokannat tarjoavat perinteisiin tietokantoihin verrattuna enemmän joustavuutta ja laajennettavuutta, parempaa tukea rakenteettomalle tiedolle, tukea helpolle kopioinnille ja ennen kaikkea tukea massiiviselle määrälle dataa. Ne ovatkin yleistyneet voimakkaasti big datan varastointiratkaisuina data-analytiikassa. (Chen et al. 2014, s. 38)

NoSQL-tietokannat voidaan jakaa karkeasti kolmeen päätyyppiin: avain-arvoparitietokannat, sarakeperhetietokannat sekä dokumenttitietokannat. Avain-arvoparitietokannoissa data yksin-

kertaisesti tallennetaan yksilöllisten avainarvojen mukaan. Jokainen avainarvo vastaa jotain dataa, ja tietokannasta voi hakea tietoa avainarvojen perusteella. Avain-arvoparitietokannat tarjoavat perinteisiin relaatiotietokantoihin nähden helpompaa laajennettavuutta ja pienempiä hakujen viiveitä. Dokumenttitietokannat varastoivat ja käsittelevät dataa dokumentteina, ja data varastoidaan osittain rakenteellisessa muodossa. Data tallennetaan nimistä ja arvoista koostuvina pareina, mutta toisin kuin avain-arvoparitietokannoissa, data on dokumenttitietokannoissa haettavissa nimen ja myös sen arvon perusteella. Relaatiotietokannoista poiketen dokumenttitietokannat sopivat erityisen hyvin big datan varastointiin, sillä ne toimivat hyvin rakenteettoman datan, kuten tekstidokumenttien tallentamiseen. Sarakeperhetietokannat puolestaan varastoivat ja käsittelevät dataa sarakkeiden perusteella rivien sijaan ja niissä sarakkeet ja rivit segmentoidaan useisiin osiin paremman laajennettavuuden saavuttamiseksi. (Chen et al. 2014, s. 38-42; Gupta et al. 2012, s. 47-48; Moniruzzaman & Hossain 2013, s. 4-7; Zhao et al. 2014, s. 30)

Mikään yksittäinen tietokantaratkaisu ei ole optimaalinen kaikkiin mahdollisiin käyttötarkoituksiin. CAP-teoreeman mukaisesti hajautetuissa tietokannoissa on mahdotonta saavuttaa samaan aikaan eheyttä, saatavuutta ja osioinnin sietokykyä, vaan yhdestä näistä ominaisuuksista joudutaan tinkimään. Lisäksi pilvilaskentaa hyödynnettäessä tietokannoissa on tehtävä kompromisseja luku- ja kirjoitusnopeuksien, viiveen ja kestävyyden, kopioinnin, segmentoinnin ja muiden toimintojen välillä. Jokainen eri tietokantaratkaisu soveltuukin parhaiten vain tietyille sovellusalueille. (Chen et al. 2014, s. 44-45)

NoSQL-tietokantojen keskeisiksi käyttökohteiksi ovat nousseet suuren mittakaavan datan prosessointi hajautetuissa järjestelmissä, rakenteettoman tiedon analysointi sekä suurten tietomäärien varastointi tilanteissa, joissa vaatimukset tietokannan eheyden suhteen eivät ole liian tiukkoja. Tämä johtuu osin siitä, että useissa NoSQL-tietokannoissa on päädytty ratkaisuun, jossa tietokannan eheyden suhteen on tingitty paremman saatavuuden ja osioinnin sietokyvyn saavuttamiseksi. (Moniruzzaman & Hossain 2013, s. 4)

NoSQL-tietokantojen vahvuudet ja heikkoudet relaatiotietokantoihin verrattuna

NoSQL-tietokantojen etuihin data-analytiikkaa tekevässä yrityksessä lukeutuu ennen kaikkea niiden soveltuvuus rakenteettoman tiedon käsittelyyn sekä suurten datamäärien analyysiin ja

varastointiin (Moniruzzaman & Hossain 2013, s. 11-12). Erityisesti massiivista datamäärää hyödyntävässä offline-analytiikassa NoSQL-ratkaisut ovat erittäin hyvä ratkaisu perinteisten tietokantojen rajoittuneisuuden vuoksi. (Chen et al. 2014, s. 56)

Vuonna 2012 valmistuneessa NoSQL:n tulevaisuudennäkymiä kartoittaneessa tutkimuksessa 1300 yritysedustajan haastattelun perusteella todettiin, että suurimmat syyt NoSQL:n käytön lisäämiselle yrityksissä ovat olleet relaatiotietokantojen ja niiden skeemojen kankeus, tietokannan horisontaalisen laajentumisen vaikeus, heikko suorituskyky sekä kustannukset. Erityisesti suuria datamääriä, sekä paljon rakenteetonta dataa käsittelevät yritykset ovat viime vuosina siirtyneet nopeasti käyttämään NoSQL-tietokantoja. (Moniruzzaman & Hossain 2013, s. 10-11)

NoSQL-tietokantojen etuihin relaatiotietokantoihin verrattuna kuuluu myös se, että niissä on hyvin paljon vaihtelua ja valinnanvaraa, ja monelle eri sovellusalueelle löytyy juuri siihen käyttötarkoitukseen optimoitu NoSQL-tietokantaratkaisu. Tämä eroaa perinteisistä SQL-tietokannoista, jotka noudattavat pitkälti periaatetta, jossa on tarjolla sama ratkaisu kaikkiin eri sovellusalueisiin. NoSQL-tietokantojen kohdalla yrityksen onkin mahdollista valita juuri omaan toimintaan tai projektiin parhaiten soveltuva NoSQL-tietokanta. (Zhao et al. 2014, s. 23)

NoSQL-tietokantojen ylläpidon on huomattu olevan helpompaa ja yrityksen kannalta halvempaa kuin perinteisten relaatiotietokantojen ylläpidon. Nykyaikaisetkin relaatiotietokantojen hallintajärjestelmät vaativat ylläpitoonsa suuren määrän kalliita ja koulutettuja ylläpitäjiä, ja relaatiotietokantojen asennus sekä optimointi vaativat paljon asiantuntijoiden työpanosta. NoSQL-tietokannat sen sijaan ovat suunniteltu erittäin helpoiksi asentaa ja ylläpitää, ja niihin on sisällytetty muun muassa paljon automaattisia korjaustoimintoja, jotka tekevät ylläpidosta helpompaa ja halvempaa yritykselle. (Zhao et al. 2014. s. 43)

NoSQL-ratkaisut tarjoavat yritykselle kustannustehokkaamman tavan toteuttaa tiedon varastointia. NoSQL-tietokannat hyödyntävät pääsääntöisesti halvoista kuluttajapalvelimista koostuvia klustereita suurien datamäärien hallintaan, kun taas perinteiset relaatiotietokannat käyttävät pääsääntöisesti kalliita erityispalvelimia. Tämän ansiosta NoSQL-tietokantoja käyttämällä yrityksen on mahdollista saavuttaa kustannussäästöjä halvempien palvelinkustannusten ja siten

datan käsittelykustannusten muodossa verrattuna perinteisiin ratkaisuihin. (Zhao et al. 2014. s. 43)

Yrityksen on mahdollista saavuttaa kustannussäästöjä NoSQL-tietokantojen avulla myös siksi, koska NoSQL-puolella on laajasti käytössä lisenssimaksuttomia ohjelmistoja palvelinten hallintaan. Nämä lisenssimaksuttomat ohjelmistot mahdollistavat tietokannan horisontaalisen skaalautumisen pienemmillä kustannuksilla. Näiden taloudellisten etujen ansiosta NoSQL-tietokannat mahdollistavat yritykselle useita kertoja edullisemman ratkaisun datan varastointiin ja käsittelyyn verrattuna perinteisiin relaatiotietokantoihin. (Zhao et al. 2014. s. 43)

Datan määrän kasvaessa nopeasti myös tarve laajentaa tietokantoja horisontaalisesti nousee yhä ajankohtaisemmaksi. NoSQL-tietokannat ovat tässä suhteessa ylivertaisia perinteisiin relaatiotietokantoihin nähden, sillä ne ovat alusta asti suunniteltuja laajentumaan horisontaalisesti hyödyntäen edullisista kuluttajapalvelimista koostuvia klustereita. Käytännössä tämä tarkoittaa sitä, että NoSQL:n mahdollistamana suuressa mittakaavassa kokonaiskapasiteettia on edullisempaa kasvattaa hankkimalla uusi rinnakkainen palvelinyksikkö, kuin tekemällä nykyisestä palvelimesta tehokkaampi. Näin ollen ne tarjoavat big data -analytiikkaa tekevälle yritykselle varmemman vaihtoehdon tulevaisuutta ajatellen monipuolisemman laajennettavuuden muodossa. (Zhao et al. 2014. s. 43)

NoSQL-tietokannat tarjoavat enemmän joustavuutta ja vähemmän rajoitteita datamallien suhteen, eli toisin sanoin datan tallennusmallin (esimerkiksi skeema) muutos on helpommin toteutettavissa NoSQL-tietokannoilla. Tämä on selvä etu verrattuna perinteisiin relaatiotietokantoihin, joissa pienetkin muutokset datamalleihin vaativat huolellista toteutusta ja usein myös palveluiden väliaikaista alasajoa, mikä voi olla yrityksen liiketoiminnalle hyvin kallista. Näin ollen NoSQL-ratkaisut sopivat hyvin sovelluksissa, joissa käytetyssä datassa on paljon vaihtelua, kuten big data -analytiikassa. (Zhao et al. 2014. s. 43)

NoSQL-tietokantojen käytössä tiedon hakuun ja analysointiin tarvitaan usein paljon ohjelmointiosaamista yksinkertaisissakin tietokantakyselyissä. Lisäksi rakenteettoman tai osittain rakenteellisen tiedon analysoinnissa NoSQL-tietokannat vaativat yritykseltä analytiikka-osaamisen lisäksi runsaasti ohjelmointiosaamista (Moniruzzaman & Hossain 2013, s. 4). Tämä on suuri

rasite etenkin pienille yrityksille, joille ohjelmointiosaamisen hankkiminen voi olla hyvin kallista. (Zhao et al. 2014. s. 44)

NoSQL-tietokannoissa on heikko tuki transaktioille, eli tietokannan siirtymiselle tilasta toiseen siten, että kaikki muutokseen kuuluvat toimenpiteet toteutuvat. Lisäksi NoSQL-tietokannoissa on usein tingitty eheyden saavuttamisesta, minkä vuoksi ne saavuttavat muutoksen jälkeen eheän tilan vasta tietyn viiveen jälkeen (Moniruzzaman & Hossain 2013, s. 4). Toisin sanoin, järjestelmän sisältämien datan kopioiden välillä voi olla jonkin aikaa eroja, eikä eri paikoissa olevat kopiot ole saman tien identtisiä. Tämä tekee NoSQL-tietokannoista vähemmän luotettavia kuin relaatiotietokannoista liiketoiminnan kannalta elintärkeiden järjestelmien tietokantoina. (Zhao et al. 2014. s. 44)

Monet NoSQL-tietokantaratkaisut ovat edelleen kehitysvaiheessa, mikä tarkoittaa sitä, että niissä voi olla puutteita toimintojen monipuolisuuden sekä luotettavuuden suhteen. Tästä syystä yritysten tulee olla varovaisia implementoidessaan NoSQL-tietokantoja tärkeisiin järjestelmiin. Suurin osa NoSQL-ratkaisuista on avoimen lähdekoodin projekteja, eli yrityksen on hyvin vaikeaa saada tukea niiden käyttöön ongelmatilanteissa. Tämäkin on yksi syy yrityksille lähestyä NoSQL-tietokantoja varovaisuudella, sillä vikatilanteiden sattuessa tuen löytäminen voi olla hidasta tai kallista. Tämä eroaa merkittävästi relaatiotietokannoista, joiden toimittajat tarjoavat asiakkailleen nopeaa ja tehokasta tukea. (Zhao et al. 2014. s. 44)

Työmarkkinoilla on tänä päivänä huomattavasti vaikeampi löytää osaavia ammattilaisia ylläpitämään NoSQL-tietokantoja verrattuna perinteisten relaatiotietokantojen ylläpitäjiin. Lisäksi NoSQL-tietokantojen kehittäjät ovat itsekin vielä paljolti oppimisvaiheessa, sillä NoSQL tekniikkana ei ole vielä kovin kypsää verrattuna vuosikymmeniä markkinoilla olleisiin relaatiotietokantoihin. Tämä voi osoittautua ongelmaksi esimerkiksi pienille toimijoille, joiden on vaikeaa rekrytoida tarvitsemiaan asiantuntijoita NoSQL-projektin tueksi. (Zhao et al. 2014. s. 44)

Tietokantaratkaisut offline- ja online -analytiikassa

Online-analytiikassa, eli reaaliaikaisesti toteutettavassa data-analytiikassa, on laajasti käytössä perinteisiä relaatiotietokantoihin perustuvia arkkitehtuuriratkaisuja, vaikkakin MongoDB:n

kaltaiset NoSQL-tietokantaratkaisut ovat yleistyneet viime vuosina. Online-analytiikassa käsitellään liikkuvaa dataa, eli dataa analysoidaan jatkuvasti sitä mukaan, kun sitä syntyy. Online-analytiikan etuihin lukeutuu juuri-oikealla-hetkellä -toimintamalli eli JIT (Just-In-Time), jonka avulla liiketoiminnan kannalta tärkeää tietoa saadaan luotua reaaliaikaisesti yrityksen johdon käyttöön. (Chen et al. 2014, s. 55-56; Gupta et al. 2012, s. 50-51)

Offline-analytiikalle on tyypillistä se, että sen avulla käsitellään massiivisia datamääriä, jotka ylittävät perinteisten relaatiotietokantojen ja analytiikkatyökalujen kapasiteetin. Offline-analytiikassa hyödynnetäänkin paljolti Hadoopin ja MapReducen kaltaisia menetelmiä sekä NoSQL-tietokantoja. (Chen et al. 2014, s. 55-56)

Yhteenveto tietokantaratkaisun valinnasta big data -analytiikassa

Tietokantoja valittaessa suurin päätös liittyy siihen, käytetäänkö perinteistä relaatiotietokantaa vai jotain uudempaa NoSQL-tietokantaratkaisua. On selvää, että yksikään yksittäinen ratkaisu ei sovi kaikkiin tilanteisiin ja käyttökohteisiin, vaan optimaalisen tietokannan valinta riippuu täysin siitä, millaiseen sovellukseen yritys on sitä tarvitsemassa. Data-analytiikan kannalta on kuitenkin ilmeistä, että NoSQL-tietokannat tarjoavat useimmissa tilanteissa selvästi paremmat työkalut big datan käsittelyyn kuin perinteiset relaatiotietokannat. (Chen et al. 2014, s. 38; Zhao et al. 2014. s. 44-45)

On tärkeä huomata, että tänä päivänä analytiikkaa tekevien yritysten edessä ei kuitenkaan ole mustavalkoista valintaa SQL-pohjaisten relaatiotietokantojen ja uusien NoSQL-tietokantojen välillä, tai että toinen tietokantamalli olisi korvaamassa toisen kokonaan (Grolinger et al. 2013, s. 19). Sen sijaan molemmat tekniikat ovat hyödynnettävissä rinnakkain, ja uudet kehitysaskeleet SQL-tietokannoissa ovatkin tuomassa relaatiotietokantoja lähemmäksi NoSQL-tietokantoja big data -analytiikassa esimerkiksi tukemalla paremmin tietokannan skaalautuvuutta. Analytiikkaa tekevän yrityksen kannalta oleellinen valinta onkin se, mitä ratkaisua kannattaa käyttää missäkin tilanteessa ilman, että mitään vaihtoehtoja rajataan kokonaan pois. (Zhao et al. 2014. s. 44-45; Vaas 2016)

Hyvin paljon tietokantaratkaisun valintaan vaikuttaa se, kuinka rakenteellista tai rakenteetonta dataa ollaan analysoimassa, ja kuinka paljon tietokantaa ollaan aikeissa skaalata suuremmaksi

tulevaisuudessa. Siinä missä perinteisillä relaatiotietokannoilla on heikko tuki rakenteettomalle tiedolle ja skaalautuvuudelle, voidaan erilaisten NoSQL-ratkaisujen avulla luoda erittäin skaalautuvia tietokantaratkaisuja, joiden avulla onnistuu tehokkaasti myös rakenteettoman datan varastointi ja analysointi. Lisäksi tietokannan eheysvaatimuksella on suuri merkitys tietokantaratkaisun valintaan. (Loshin 2013, s. 113)

Oikean ratkaisun valintaan vaikuttaa paljolti se, millaisesta yrityksestä ja toimialasta on kyse, sekä millaista dataa ollaan analysoimassa. Esimerkiksi finanssialalla ja verkkokaupassa tallennettu asiakasdata on hyvin rakenteellista, ja maksujärjestelmien toimintavarmuuden vuoksi käytössä on pitkälti korkean eheyden takaavia relaatiotietokantoja. Lisäksi näillä aloilla on yleisesti käytössä reaaliaikaista datan käsittelyä ja analysointia. Tämänkaltaisissa tilanteissa erilaiset SQL-pohjaiset tietokantaratkaisut ovat yleinen ja luontevin valinta. Oheisessa taulukossa 1 on havainnollistettu relaatiotietokantojen ja NoSQL-tietokantojen keskeisiä eroja, joita tämän työn aikana on nostettu esille. (Chen et al. 2014, s. 55-56)

Taulukko 1 Relaatiotietokantojen ja NoSQL-tietokantojen keskeisiä eroja (mukaillen Levin 2015)

	Relaatiotietokanta	NoSQL
Datan tyyppi	Rakenteellinen	Ei-rakenteellinen
Sopisi massiiviseen	Taulukkolaskentatiedostoon	Tekstitiedostoon
Skeema	Pysyy samana	Muuttuu usein
Yhteensopivat datamuodot	Käyttäjätiedot, varastotiedot	Kuvat, videot, viestit
Yhteensopivat kyselyt	SQL	MapReduce, Python
Transaktiotuki	Korkea	Matala
Monipuolisuus	Yksi ratkaisu kaikkiin tarpeisiin	Eri käyttökohteisiin oma optimaalinen ratkaisu
Hankinta ja tuki	Toimittajia paljon ja tukea saa nopeasti	Toimittajia niukasti ja tuen saaminen hankalampaa
Kustannukset	Hankinta- ja ylläpitokustannukset suuria	Kustannukset pienempiä

Taulukosta 1 nähdään, että relaatiotietokannat eivät sovi rakenteettomaan dataan, jollaista big data tyypillisesti on. Lisäksi tiukka skeemarakenne tekee relaatiotietokannoista heikosti skaalautuvia, mikä on big data -analytiikassa oleellista datan varastoinnin näkökulmasta. Tämän vuoksi NoSQL-tietokannat ovat useimmissa tilanteissa sopivampi ratkaisu datan varastointiin big data -analytiikassa. Toisaalta NoSQL-tietokannoissa on usein joustettu eheyden suhteen eikä niillä ole pääsääntöisesti hyvää tukea transaktioille, toisin kuin relaatiotietokantojen kohdalla.

Taulukossa 1 ilmenee myös se, että NoSQL-tietokannoissa on tarjolla erilaisiin käyttötarkoitukseen runsaasti erilaisia ratkaisuja, joista jokainen on optimaalinen juuri tiettyyn käyttötarkoitukseen. Sen sijaan relaatiotietokannoissa monipuolisuus on rajallisempaa ja eri data-analytiikan käyttötarkoituksiin ei ole saatavilla yhtä hyvin optimoituja tietokantaratkaisuja. Lisäksi relaatiotietokantojen hankinta ja ylläpito ovat pääsääntöisesti kalliimpia kuin NoSQL-tietokantojen kohdalla. Relaatiotietokannoissa on kuitenkin saatavilla enemmän toimittajia markkinoilla, mikä tekee niiden hankinnasta ja tuen saamisesta niille helpompaa.

Tietokannan rakenteen lisäksi data-analytiikassa on tärkeää tehdä päätös sen suhteen, alkaako yritys käyttää ulkoistettua datan varastointiratkaisua kuten erilaiset pilvipalvelut, vai käytetäänkö analytiikkaan yrityksen omissa tiloissa sijaitsevaa fyysistä infrastruktuuria. Pilvipalvelun valintaa puoltavia tekijöitä voivat olla muun muassa kustannustehokkuus, sitoutuvan pääoman pienentäminen, käyttöönoton nopeus ja helppous sekä palvelun kustannusten joustavuus käyttömäärän suhteen. Oman infrastruktuurin hankintaa puoltavia tekijöitä taas ovat muun muassa tietoturva, pienemmät viiveet, saatavuus sekä toimintavarmuus. Näitä asioita käsitellään laajemmin seuraavassa kappaleessa. (Zhao et al. 2014, s. 5-6, 40-41)

4 PILVIPALVELUT

4.1 Pilvipalvelut data-analytiikassa

Pilvipalvelu data-analytiikassa tarkoittaa ulkoisilta kumppaneilta ostetun tallennus- ja laskentakapasiteetin käyttämistä oman datan varastointiin ja prosessointiin. Pilvilaskentaa käytetään suuren mittakaavan data-analytiikan toteuttamiseen skaalautuvasti ja kustannustehokkaasti. Pilvipalveluiden määritelmiä on monia, mutta niille yhteistä on se, että ne erottelevat pilvipalveluissa seuraavat ominaisuudet: palvelu on tarjolla automaattisesti asiakkaan tarpeen mukaan, palvelu on saatavilla laajasti erilaisilla laitteilla verkon yli, palveluntarjoajan laskentaresurssit on yhdistetty palvelemaan samaan aikaan useita asiakkaita, palvelun tarjoama kapasiteetti skaalautuu asiakkaan tarpeen mukaan sekä palvelun käytön määrä on mitattavissa ja raportoitavissa jokaisen asiakkaan kohdalla (Zhao et al. 2014, s. 10-11). (Gupta et al. 2012, s. 42, 44-45)

Pilvipalveluiden hyödyntäminen IT-infrastruktuuriratkaisuna on yleistynyt merkittävästi viime vuosina. IT-infrastruktuurin tarjoamisen ja ylläpidon ulkoistaminen kumppaneille on nähty tarjoavan kustannussäästöjä IT-resurssien hallintaan liittyen. Tämä perustuu siihen, että pilvipalveluiden tarjoajat saavuttavat mittakaavaetuja tarjotessaan kapasiteettiaan suurelle määrälle asiakkaita, mikä mahdollistaa kapasiteetin tarjoamisen aikaisempaa kustannustehokkaammin. Toisaalta pilvipalvelut tarjoavat myös asiakasyrityksille mahdollisuuden vastaanottaa tietokonekapasiteettia palvelun muodossa joustavasti tarpeen mukaan, mikä lisää tämän ratkaisun houkuttavuutta verrattuna oman IT-infrastruktuurin hankintaan ja ylläpitoon. (Gupta et al. 2012, s. 42, 44-45; Zhao et al. 2014, s. 3)

Pilvipalvelut eivät ole nopeasti kasvaneesta suosiostaan huolimatta kovinkaan uusi tai vallankumouksellinen tekniikka, vaan ne ovat jo pitkään kestäneen tietotekniikan kehityksen tulos. Jo viime vuosisadalla oli olemassa tietokonekeskuksia, joita useat yritykset pystyivät käyttämään yhteistyössä saadakseen tarvitsemaansa IT-kapasiteettia. Nykyaikaiset pilvipalvelut ovatkin kehittyneet tämän ajatuksen pohjalta, mutta niiden toteuttaminen nykyisessä muodossaan on vaatinut paljon kehitystä muun muassa automaattisen tietojenkäsittelyn, tietoliikenneverkkojen ja virtualisaation saralla. Viime vuosina etenkin big datan yleistyminen on luonut uusia

haasteita ja mahdollisuuksia pilvipalveluiden käytölle datan varastoinnin näkökulmasta esimerkiksi kasvaneiden tallennuskapasiteettia ja laskentatehoa koskevien vaatimusten kautta. (Gupta et al. 2012, s. 43; Zhao et al. 2014, s. 3-4)

Pilvilaskennan ja pilvipalveluiden onnistumiseen vaikuttaa useita tekijöitä, kuten palvelun saatavuus, skaalautuvuus, muokkautuvuus, suorituskyky ja vikatilojen sietokyky. Palvelun toimivuuden ja menestyksen kannalta on oleellista, että palvelu on saatavilla kaikkina aikoina, vaikka esimerkiksi verkossa olisi ongelmia tai jokin datakeskus poistuisi verkosta. Palvelun on oltava skaalautuva ja muokkautuva, eli kapasiteettia on oltava tarjolla asiakkaan tarpeen mukainen määrä, ja palvelun on mukauduttava esimerkiksi asiakkaan tietokantahakutarpeiden mukaan. Koska palvelun hinnoittelu riippuu käytetystä kapasiteetista, on toiminnan kannattavuuden kannalta myös suorituskyvyllä erittäin suuri merkitys. Mikä tärkeintä pilvipalvelun on kyettävä sietämään vikatiloja ilman, että esimerkiksi menetetään kokonaan dataa. (Gupta et al. 2012, s. 44-45; Zhao et al. 2014, s. 4-5)

Pilvipalveluiden yleistymiselle on olemassa lukuisia haasteita, jotka voivat tulevaisuudessa rajoittaa tämän palvelutyypin yleistymistä. Suurin haaste liittyy palvelun saatavuuteen, sillä se on asiakkaiden näkökulmasta erittäin kriittinen tekijä palvelun käytettävyyden kannalta. Lyhyetkin katkot palvelun toiminnassa voivat olla hyvin kalliita asiakasyrityksille, minkä vuoksi monet yritykset ovatkin huolissaan pilvipalveluiden luotettavuudesta. (Gupta et al. 2012, s. 45; Zhao et al. 2014, s. 5)

Toinen keskeinen pilvipalveluiden yleistymisen haaste liittyy tietoturvaan (Januzaj et al. 2015, s. 1857). Datan varastoiminen pilvipalveluihin yrityksen omien palvelimien sijaan on aina riskialtista tietoturvan näkökulmasta, sillä heikosti salattu data voi päätyä kolmannen osapuolen luettavaksi. Tämä on erityisen riskialtista yrityksen liiketoimintaan oleellisesti liittyvän datan tai esimerkiksi asiakastietoja sisältävän datan kohdalla. (Gupta et al. 2012, s. 45; Zhao et al. 2014, s. 5)

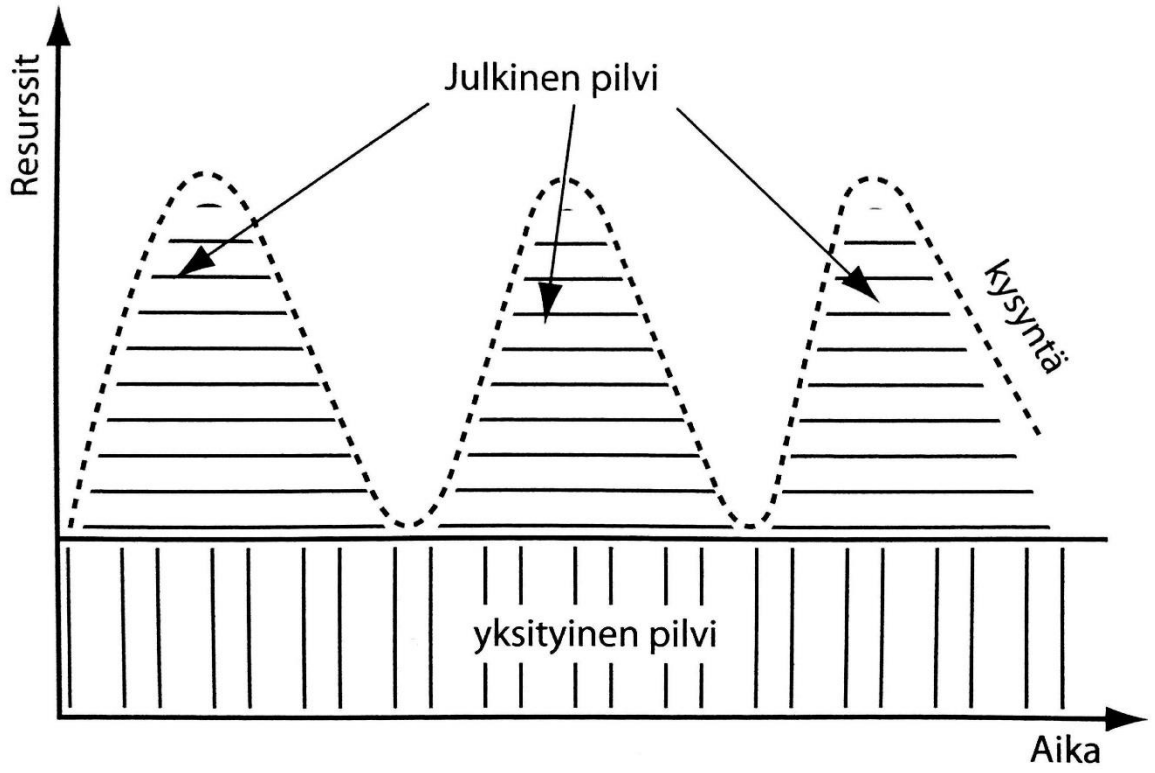
Datan lukkiintuneisuus on myös yksi pilvipalveluiden haaste. Tällä tarkoitetaan sitä, että pilvipalveluiden ohjelmointirajapintoja ei ole vielä täysin standardisoitu, eikä tämän vuoksi datan siirtäminen yhdestä pilvipalvelusta toiseen ole aina mahdollista. Tämä on monille yrityksille

yksi este pilvipalveluiden laajemmalle käyttöönotolle. (Salo 2010, s. 114; Zhao et al. 2014, s. 5)

4.2 Pilvipalveluiden tekniset ratkaisut

Pilvipalvelut voidaan jakaa kolmeen yleisimpään alaluokkaan sen perusteella, millä tavalla käyttäjälle eli asiakkaalle varataan ja toimitetaan myytävät laskentaresurssit. Niitä ovat yksityinen, julkinen ja hybridi-pilvi. Yksityisellä pilvellä tarkoitetaan, että ostettavat resurssit ovat myös fyysisesti varattu sen ostaman asiakkaan käyttöön mahdollistaen pilven resurssien hyödyntämisen yrityksen eri toimipaikoista. Suoritustehon tehokas jakaminen ja rajaaminen työkuormille tehdään virtualisointiteknologiaa hyödyntäen. Yksityisen pilven etuina ovat korkea tietoturvan ja datan yksityisyyden hallinnan taso, sillä pääsy palvelimille, joilla dataa säilytetään, on ainoastaan yrityksellä itsellään palomuurin takana. Koska yksityistä pilveä tarjoava yritys ei pysty jakamaan resursseja yhtä tehokkaasti laajan käyttäjäjoukon kanssa, ovat myös sen kustannukset muita vaihtoehtoja korkeammat. (Salo 2010, s. 32)

Julkinen pilvi puolestaan käsittää palvelut, joissa massiivisia datakeskuksia omistavat yritykset myyvät laskentatehoa käyttäjille (Salo 2010, s. 34-35). Erona on, että samoja palvelimia käyttävät useat eri asiakkaat samanaikaisesti, vaikka ne ovatkin virtuaalikonein rajattuina. Suuri julkisen pilven hyöty verrattuna yksityiseen on mahdollisuus yksittäiselle asiakkaalle huolettomasti skaalata tarvittavaa resurssien määrää niin ylös kuin alas, sekä maksaa ainoastaan käytetystä laskentatehosta. Kolmas ratkaisuvaihtoehto on toteuttaa hybridipilvi, eli julkisen ja yksityisen pilven yhdistelmäratkaisu. Liiketoiminnan kannalta kriittisimmät työkuormat pystytään tämän tekniikan avulla tällöin suorittamaan luotettavasti ja turvallisesti yksityisen pilven avulla, mutta samalla julkisen pilven loputon laskentateho pystytään hyödyntämään raskaimpien laskutoimitusten toteuttamiseen. (Assunção et al. 2015, s. 6; Zhao et al. 2014, s. 15-16)



Kuva 3 Muuttuvaan resurssitarpeeseen vastaaminen hybridipilviratkaisulla (Salo 2010, s. 89)

Yllä olevan kuvan 3 mukaisesti hybridipilvi voi palvella yrityksen resurssitarvetta tarjoamalla käyttöön helposti mukautettavan määrän laskentatehoa. Liiketoiminnalle kriittiset sovellukset ja arkaluonteinen tai kallisarvoinen data pidetään yrityksen yksityisessä pilvessä, joka mitoitetaan minimikapasiteetin mukaisesti. Vaihtuvaan resurssitarpeeseen vastataan ottamalla tehoja käyttöön julkisesta pilvestä. Se, kuinka suureksi yksityisen pilven kapasiteetin mitoittaa, riippuu paljon yrityksen koosta, sillä se vaatii ylläpitoa ja sitoutuneita kustannuksia enemmän kuin julkinen pilvi. Toisaalta jos yritys on todella suuri, on sillä enemmän IT-toimintoja jotka ovat kriittisiä liiketoiminnalle, jolloin yksityisen pilven osuus voi olla myös suurempi. (Salo 2010, s. 89)

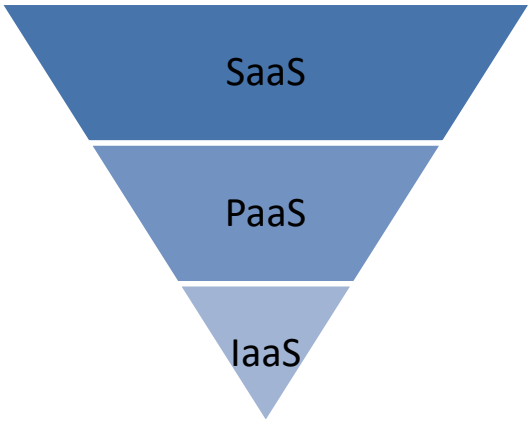
4.3 Pilven palvelu- ja liiketoimintamallit

Tarjolla olevia pilvi- eli etäresurssipalveluita on markkinoilla hyvin monenlaisia eri variaatioita. Pilvipalvelua hankittaessa tarkasteltavaksi tulee erityisesti se, kuinka paljon haluaa itse pystyä vaikuttamaan lopullisen ratkaisukokonaisuuden ominaisuuksiin ja verrata sitä siihen,

kuinka suuri hyöty saadaan, kun vastuuta ylläpidosta ja toiminnasta voidaan siirtää pois yrityksen sisältä. Pilvipalveluratkaisuja tarjotaan erilaisin liiketoimintamallein, joita voidaan vertailla niiden tarjoaman mukautettavuuden ja käytön huolettomuuden suhteen. (Heino 2010, s. 50-54; Salo 2010, s. 22-30)

Matalimman tason pilvipalveluratkaisu on infrastruktuuri palveluna (Infrastructure-as-a-Service, IaaS). Tässä liiketoimintamallissa palvelun ylläpitäjä tarjoaa asiakkailleen virtuaalista konesalikapasiteettia. Käytännössä IaaS tarjoaa käyttäjän tarvitsemat kapasiteettiresurssit palveluna korvaamaan yrityksen omat fyysiset laitteistot, mutta kaikki ohjelmisto mukaan lukien käyttöjärjestelmät jäävät asiakkaan vastuulle. (Heino 2010, s. 50-54; Salo 2010, s. 22-30)

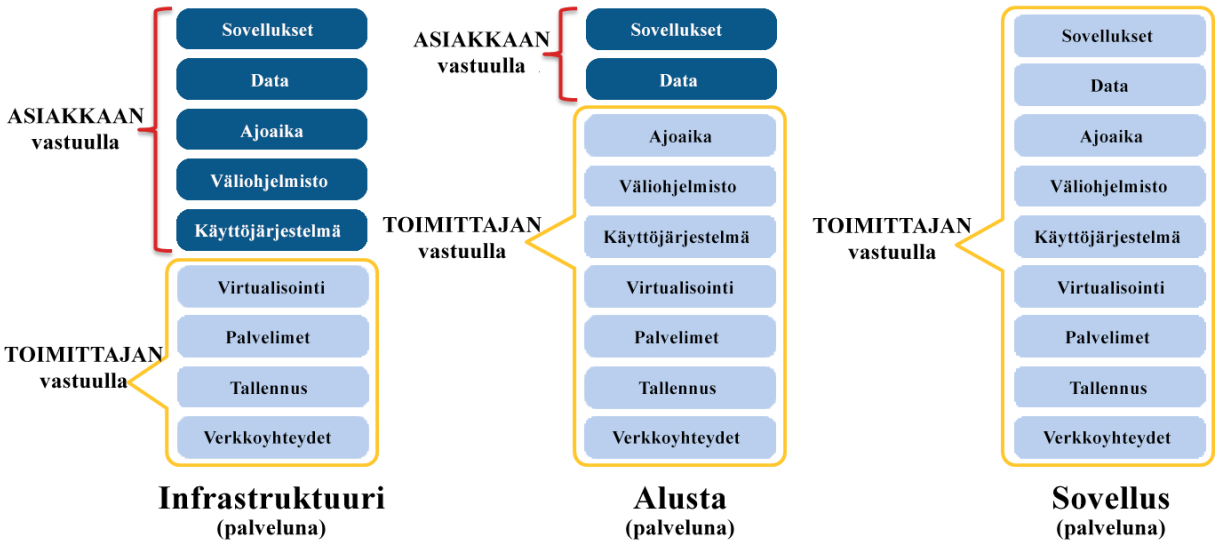
Hieman IaaS-mallia pidemmälle viety vaihtoehto on kuvan 4 mukaisesti hankkia sovellusalusta palveluna (Platform-as-a-Service, PaaS), jolloin ratkaisuntarjoajan vastuulle kuuluu edellisten lisäksi myös alustan toimimiseen vaadittavat ominaisuudet, kuten käyttöjärjestelmä ja laiteajurit. PaaS-palvelumallissa asiakas itse vastaa ohjelmien ja datan käsittelystä, mutta toisaalta pystyy myös kustomoimaan niitä vastaamaan paremmin omia tarpeitansa. Käytännössä rajoittaviksi tekijöiksi voivat muodostua rajattu tuki eri ohjelmointikielille tai rajapinnoille. (Heino 2010, s. 50-54; Salo 2010, s. 22-30)



Kuva 4 Pilvipalveluiden liiketoimintamallit (Zhao et al. 2014, s.14)

Korkein pilvipalvelumallin taso on etäsovelluspalvelu (Software-as-a-Service, SaaS), jolloin koko asiakkaan tarvitsema ohjelmisto tarjotaan sellaisenaan ja käyttövalmiina verkkoyhteyden

välityksellä. Tällöin asiakkaan ei tarvitse huolehtia lainkaan ratkaisukokonaisuuteen liittyvistä teknisistä taustatoiminnoista. Kuvassa 5 on havainnollistettu eri liiketoimintamalleihin liittyviä vastuualueita. (Heino 2010, s. 50-54; Salo 2010, s. 22-30; Zhao et al. 2014, s.14-15)



Kuva 5 Asiakkaan ja toimittajan vastuualueet pilvipalveluiden liiketoimintamalleissa (Romanek 2015)

Analytiikan käyttöön on luotu myös spesifimpiä ratkaisuja perinteisten SaaS-, PaaS-, ja IaaS-ratkaisujen rinnalle. Tietokanta palveluna (Database-as-a-service, DaaS tai DBaaS) on palvelu, jossa yritys tarjoaa pilvipalvelun muodossa tietokantoja muiden yritysten käyttöön (Boukhelef et al. 2016, s. 1). Tämä palvelu tarjoaa asiakasyrityksille mahdollisuuden ostaa juuri optimaalinen tietokanta käyttöönsä ilman laiteinvestointeja, asiantuntijoiden rekrytointeja ja ylläpitokustannuksia. (Zhao et al. 2014, s. 32-33)

Pay-per-use on liiketoimintamalli, jossa asiakas maksaa saamastaan palvelusta vain todellisen käytön määrän mukaan. Pilvipalveluissa tämä tarkoittaa esimerkiksi sitä, että asiakas maksaa vain käyttämänsä laskenta- tai tallennuskapasiteetin verran. Pay-per-use palveluita tarjoavat yritykset pyrkivät saamaan laskentatehon hinnan matalaksi sillä, että niiden valtavien konesalien käyttöaste pysyy korkeana ja hajonta pienenä useiden samanaikaisten asiakkaiden ansiosta. Pay-per-use -mallin vahvuuksiin kuuluu siis kustannusten skaalautuvuus palvelun käyttömäärän mukaan sekä se, että tässä mallissa asiakas ei joudu tekemään esimerkiksi suuria käyttöönottoinvestointeja. (Zhao et al. 2014, s. 9-20)

Yrityksen kustannusrakennetta tarkastellessa laskentatoimen näkökulmasta eroavaisuus IT:stä aiheutuvista välittömistä kustannuksista näkyy kiinteiden kustannusten poistumisena pilvipalvelumallia käytettäessä. Ostettaessa järjestelmät palveluna ei niihin sitoudu yrityksen pääomaa, eikä käyttämättömästä kapasiteetista joudu myöskään maksamaan. Lisäksi huollon ja ylläpidon kustannukset kuuluvat ennakkoon tiedossa olevaan laskutushintaan. Toinen pilvipalvelumallin etu laskentatoimen näkökulmasta on se, että IT:stä aiheutuvat kustannukset on huomattavasti helpompi kohdentaa. Käyttöperusteinen laskutus mahdollistaa sen, että kustannuksen käytöllään aiheuttanut yrityksen osasto ja sen toiminto saadaan tietoon, toisin kuin On-Premise ratkaisuissa, jolloin kiinteiden kustannusten jakamiseen tarvittava kustannusajuri on hankala määrittää. (Salo 2010, s. 86-96)

4.4 Pilvipalveluiden valinta

Valittaessa analytiikan työkalujen ja resurssien hankintatapaa on otettava huomioon vaihtoehtojen elinkaarikustannukset, sillä vaihtoehtojen kustannusrakenteet poikkeavat huomattavasti. Muutokset tukitoimintoihin vaadittavaan työmäärään, sekä yleiseen tuottavuuden tehokkuuteen ovat myös tärkeitä kriteerejä, joita tarkastella toteutusta suunnitellessa. (Salo 2010, s. 70-77)

IT-infrastruktuurin omistaminen paikallisesti tarkoittaa sitä, että sen ylläpitoon ja huoltoon täytyy yrityksen sisältä löytyä osaavia työntekijöitä. Laitteiston ja ohjelmistojen hankinta- ja päivityskustannukset ovat paikallisen ratkaisun omistajan harteilla. IT-hankinnat ovat usein pitkäkestoinen prosessi, joka voi sitouttaa useita henkilöitä jopa vuoden ajaksi. Laitteistojen ja ohjelmistojen keskimääräinen ikä on kolmesta viiteen vuotta (Heino 2010, s. 181), sekä tarpeen vaihdellessa voi hankintaprojekteja syntyä jopa parin vuoden välein. (Salo 2010, s. 70-77)

Pilvipalveluiden edut

Pilvipalvelun etuna on, että parhaimmillaan sen käyttöönotto ei vaadi ollenkaan tuotteiden asentamista ja kaikki tarvittavat tiedostot ovat saatavilla miltä tahansa laitteelta, jolla voidaan muodostaa yhteys Internetiin (Biswas et al. 2013, s. 37). Resurssien ostaminen palveluna vapauttaa yrityksen myös päivittämiseltä ja tuelta, jota tarvittavan laitteiston ja ohjelmiston omistaminen vaatii. (Zhao et al. 2014, s. 19-20)

Yksi keskeisimmistä pilvipalveluiden käytön eduista koskee pieniä yrityksiä, joilla ei ole resursseja rakentaa riittävää IT-infrastruktuuria yrityksen sisällä. Pilvipalveluiden käyttö nimittäin pienentää merkittävästi yrityksen koon vaikutusta sen taloudelliseen menestykseen, eli se antaa pienillekin yrityksille mahdollisuuden hyödyntää laajasti monipuolisia datan varastointipalveluita, vaikka yrityksellä ei olisi resursseja rakentaa ja ylläpitää itse tarvittavaa infrastruktuuria. Pienen yrityksen ei siis tarvitse tehdä mittavia laitteistoinvestointeja luodakseen itselleen edellytyksen menestyä isompia kilpailijoita vastaan, vaan liiketoiminnan kannalta tarvittavan IT-infrastruktuurin saa hankittua käytön määrän mukaan hinnoiteltuna kustannustehokkaasti. (Salo 2010, s. 97; Zhao et al. 2014, s. 3)

Pilvipalveluiden käytön etuihin lukeutuu myös se, että pilven kautta hankittu IT-infrastruktuuri skaalautuu käytön mukaan, eli yritys saa hankittua juuri sen verran kapasiteettia, mitä se kulloinkin tarvitsee ylläpitääkseen palveluitaan. Vastaavasti myös kustannukset skaalautuvan pilvipalvelun käytöstä muodostuvat käytön määrän mukaan (pay-per-use -liiketoimintamalli). (Gupta et al. 2012, s. 44-45; Zhao et al. 2014, s. 19-20)

Pilvipalveluiden käyttö data-analytiikassa tarjoaa myös hyvät mahdollisuudet rinnakkaisajoon (parallelization), eli niiden avulla on mahdollista ajaa useita saman suuren prosessin eri instansseja rinnakkaisesti, jolloin koko prosessin suoritusaika voi lyhentyä murto-osaan. Ilman rinnakkaisajoa raskaan analytiikka-prosessin ajaminen esimerkiksi yhdellä koneella voisi vielä useita kertoja pitempään. (Zhao et al. 2014, s. 20)

Pilvipalveluiden heikkoudet

Pilvipalveluiden heikkouksiin lukeutuvat ennen kaikkea tietoturvariskit sekä palvelun saatavuuteen ja datan siirrettävyyteen pois palvelusta liittyvät huolenaiheet. Pilvipalveluiden käyttö sisällöltään hyvin arkaluonteisen datan käsittelyssä ei aina ole viisasta, sillä pilvipalveluiden kohdalla yrityksen on vaikeampi varmistaa, ettei esimerkiksi arkaluonteista asiakasdataa, kuten potilastietoja tai tilitietoja, pääse kolmannen osapuolen nähtäville. (Gupta et al. 2012, s. 45; Zhao et al. 2014, s. 5-6)

Pilvipalveluiden saatavuudessa voi esiintyä katkoksia, mikä voi olla yritykselle hyvin kallista, jos yrityksen liiketoiminnan jatkuvuuden kannalta oleellinen data ei ole tarvittaessa saatavilla

tällaisen katkoksen vuoksi. Lisäksi jos yritykselle tulee joskus tarve vaihtaa palveluntarjoajaa esimerkiksi hinnanmuutosten vuoksi, tai palveluntarjoajan toiminta lakkaa, voi datan poistaminen pilvipalvelusta ja siirtäminen toiseen palveluun olla datan lukkiintuneisuuden vuoksi hyvin vaikeaa. Nämä huolenaiheet tuleekin ottaa huomioon, kun yritys harkitsee datan varastoinnin ulkoistamista data-analytiikassa. (Gupta et al. 2012, s. 44-45; Zhao et al. 2014, s. 5-6)

Pilvessä toimivat tietokannat häviävät yrityksen omistamille fyysisille tietokannoille latenssissa, eli ajassa, mikä menee tiedon hakemisesta tiedon saamiseen tietokannasta. Lisäksi pilvessä olevien tietokantojen kohdalla myös tiedonsiirto voi olla kallista. Tästä syystä yrityksille, jotka tarvitsevat tietokannoilta korkeaa varmuutta eheyden ja saatavuuden suhteen, jotka tarvitsevat muutamissa mikrosekunneissa mitattavaa latenssia tai joiden tietokantojen käyttö on hyvin kirjoituspainotteista, pilvipalveluiden kautta toimivat tietokannat eivät ole aina hyvä ratkaisu. Tällaisia yrityksiä ovat esimerkiksi finanssialan, osakemarkkinoiden sekä toimitusketjujen hallinnan toimijat. (Zhao et al. 2014, s. 40)

Pilvipalveluna toimivissa tietokannoissa joudutaan tekemän kompromisseja latenssin sekä kestävyyden välillä. Tämä tarkoittaa sitä, että tietokantaan kirjoitus voidaan synkronoida joko ennen kuin käyttäjälle ilmoitetaan onnistuneesta kirjoittamisesta, tai vasta jälkikäteen. Jälkimmäinen ratkaisu pienentää latenssia, mutta voi johtaa tiedon häviämiseen, jos palvelin esimerkiksi kaatuu käyttäjälle raportoinnin ja tiedon tietokantaan synkronoinnin välissä. Tämä on riskialtista korkeaa tiedon varmuutta vaativille yrityksille. (Zhao et al. 2014, s. 41)

Tietokannan eri osien synkronoinnissa joudutaan myös tekemään kompromisseja sen suhteen, kopioidaanko tietoa kaikkiin eri osiin samanaikaisesti vai ei. Jos tietoa ei kopioida eri palvelimille samanaikaisesti, voidaan pienentää järjestelmän latenssia esimerkiksi tilanteissa, joissa osa palvelimista on juuri sillä hetkellä offline-tilassa. Tällöin syntyy kuitenkin riski siitä, että osa tiedon kopioista ei ole tietokannassa ajan tasalla, tai että päivitys jää kokonaan tekemättä, jos palvelin kaatuu ennen kuin siltä ehditään kopioida tietoa eteenpäin. Tämäkin on ongelmallista käyttökohteissa, joissa tietokannan suhteen on korkeat eheysvaatimukset. (Zhao et al. 2014, s. 41)

Pilvipalveluiden käyttöönoton haasteet

Pilvipalveluihin siirtymisen esteitä ja huolenaiheita yrityksissä ovat Salon (2010) mukaan erityisesti turvallisuus, saatavuus ja suorituskyky. Käyttöönoton kannalta suurimpia haasteita ovat nykyinen IT-infrastruktuuri ja siihen sitoutunut osaaminen, standardien puute pilvipalveluissa, integroitavuusmahdollisuudet omiin IT-järjestelmiin ja liian vähäiset kustomointimahdollisuudet pilvipalvelussa. (Salo 2010, s. 100-103)

Nykyisen IT-infrastruktuurin painolasti näkyy yrityksessä konkreettisesti siihen jo uponneina kustannuksina, joihin ei enää voida vaikuttaa. Pilvipalveluun siirtymisen kynnyksellä osaaminen nykyiseistä järjestelmistä on kehittynyt hyväksi, kun taas tietämys ja osaaminen pilvipalvelun käytännön operoimiseen vaadittavista asioista on olematonta. Laitteistojen ja ohjelmistojen välinen vertailu pilveä harkittaessa on suoraviivaista, kun taas monimutkaisuutta kustannusten näkökulmasta synnyttävät vaikeasti ennakoitavissa olevat kustannukset jotka johtuvat esimerkiksi osaamattomuudesta johtuvista virhetilanteista. Standardien uupuminen tekee pilvipalveluvaihtoehtojen vertailun eri vaihtoehtojen välillä hankalaa. Toiminnan siirtämisen lopulliset kustannukset alustasta toiseen on hankalasti ennustettavissa. (Salo 2010, s. 100-103)

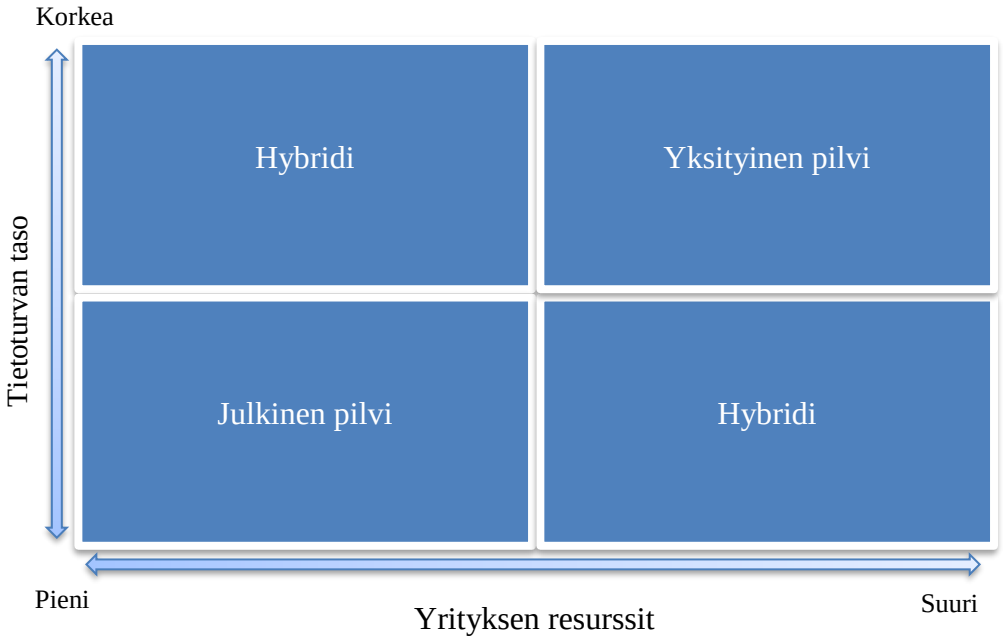
Siirtymä nykyisestä IT-ympäristöstä pilvipalveluun voi asettaa merkittäviä rajoitteita pilvipalvelun valinnalle. Datan muokkaaminen pilvipalveluun sopivaan muotoon tai sellaisen pilvipalvelun löytäminen, joka ei vaadi merkittävää vaivaa integroitumisen onnistumiseksi nykyisiin järjestelmiin ei ole nykyisistä julkisen pilven vaihtoehdoista helppoa. Erityisesti SaaS-palveluita koskee myös se, etteivät ne täytä täysin yrityksellä olevia yksilöllisiä tarpeita ohjelmiston suhteen. Niiden kustomointimahdollisuudet ovat todella vähäisiä, jolloin käyttäjä voi joutua tyytymään ainoastaan tarjolla oleviin toiminnallisuuksiin. Enemmän vapausasteita muokattavuuteen löytyy IaaS- ja PaaS-ratkaisuista, mutta tällöin saatu hyöty ylläpitotarpeen vähenemisessä pienenee. IT-järjestelmän rooli yrityksen näkökulmasta on olla liiketoiminnan funktiota ja sen tarpeita edesauttava tukitoiminto. Rajallisuudet ohjelmiston toiminnallisuuksissa juuri liiketoiminnan tarpeiden näkökulmasta voivat heijastua potentiaalista tasoa alempana suoriutumisena ohjelmiston käyttäjän työskentelyssä (Heino 2010, s. 174-175). (Salo 2010, s. 100-103)

Julkiset pilvet (public cloud) eivät aina sovellu optimaalisesti erittäin suuriin datamääriin, sillä suurten datamäärien siirtäminen pilveen voi olla hidasta ja julkisten pilvipalveluiden käyttöliittymät ovat usein liian rajoittuneita suurten datamäärien lisäämisessä. Lisäksi monilla julkisilla pilvipalveluilla on havaittu skaalautuvuusongelmia suurten datamäärien kohdalla. Tämän vuoksi yrityksen tulee pilvipalvelua etsiessään varmistaa, että tarjolla oleva palvelu tukee käyttöliittymältään ja skaalautuvuudeltaan yrityksen data-analytiikkatarpeita. (Zhao et al. 2014, s. 42)

Yhteenveto pilvipalveluiden käyttöönottoon vaikuttavista tekijöistä

Pilvipalveluiden käyttöönottoon data-analytiikassa ja datan varastoinnissa vaikuttaa ensisijaisesti palvelun luotettavuus sekä datan prosessointikustannusten suhde datan siirtokustannuksiin verkkoyhteyden kautta pilveen. Palvelun on oltava käyttö- ja datansiirtokustannuksiltaan yritykselle edullisempi, kuin datan varastoinnin toteuttaminen yrityksen oman fyysisen varastointi-infrastruktuurin avulla. Lisäksi pilvipalvelun on tarjottava yrityksen tarpeisiin nähden riittävän hyvää saatavuutta, tiedon eheyttä ja tietoturvaa. Tästä syystä pilvipalveluiden käyttöönoton järkevyyteen data-analytiikassa vaikuttaa paljolti se, millaisesta yrityksestä ja projektista on kyse. (Zhao et al. 2014, s. 41-42)

Käsiteltäessä tietoturvan kannalta hyvin kriittistä dataa, kuten asiakkaiden maksutietoja tai esimerkiksi terveystietoja, voi pilvipalveluiden käyttö datan varastointiin olla riskialtista tai jopa lailla rajoitettua tietoturvan näkökulmasta. Tällöin yrityksen omat fyysiset tietokantaratkaisut voivat olla liiketoiminnan kannalta parempi valinta kuin pilvipalvelut. Esimerkiksi finanssialan, verkkokaupan sekä sosiaali- ja terveysalan toimijoille fyysisten palvelimien ja tietokantojen hankkiminen pilvipalveluiden sijaan datan varastointiin ja analysointiin on usein paras valinta. Tällöin käyttöasteen optimoimiseksi ja kustannusten kohdentamiseksi on hyvä käyttää esimerkiksi yksityistä pilveä, johon on implementoitu tehokas virtualisointiratkaisu. (Salo 2010, s. 32)



Kuva 6 Pilvipalveluratkaisun valinta datan varastointiin big data -analytiikassa

Edellisessä kuvassa 6 on yksinkertaisella tavalla havainnollistettu keskeisiä työssä esille nousseita eroja eri pilvipalveluiden mallien välillä yrityksen ominaisuuksista ja tarpeista riippuen. Kuvassa on esitetty, kuinka yrityksen koko ja resurssit sekä tarpeet tietoturvan suhteen vaikuttavat siihen, onko yrityksen kannattavaa harkita ensisijaisesti julkista, yksityistä vai hybridipilveä. Kuvasta nähdään, että yksityinen pilvi on järkevä valinta suurille yrityksille, joilla on korkeat vaatimukset tietoturvan suhteen ja paljon resursseja käytössä. Hybridipilvi on järkevä valinta esimerkiksi pk-yrityksille, joilla on korkeita vaatimuksia datan tietoturvan suhteen, ja vastaavasti se on hyvä valinta myös suurille yrityksille, joilla tietoturvaan kohdistuvat vaatimukset eivät ole riittävän tiukkoja oikeuttamaan kalliimman yksityisen pilven hankintaa. Julkinen pilvi on hyvä valinta esimerkiksi pienelle tai keskisuurelle yritykselle, jolla ei ole tiukkoja vaatimuksia pilveen siirrettävän datan tietoturvalle ja sen resurssit eivät riitä hybridipilven käyttöönottoon. (Assunção et al. 2015, s. 5-7)

5 JOHTOPÄÄTÖKSET

Työssä käsiteltiin erilaisia datan varastointiratkaisuja ja niiden hankintakeinoja data-analytiikkaa tekevän yrityksen näkökulmasta. Työn tuloksena syntyi nelikenttämatriisi ja taulukko, joilla kuvataan graafisesti keskeisimpiä datan varastointiratkaisuun vaikuttavia tekijöitä erilaisten yritysten näkökulmasta. Nämä graafiset esitykset kiteyttävät työssä käsitellyt näkökulmat ja vertailun eri ratkaisujen välillä. Syntyneen taulukon ja nelikenttämatriisin sekä työn johtopäätösten avulla on mahdollista saada yleiskuva siitä, mitä eri tekijöitä liittyy datan varastointiratkaisun valintaan big data -analytiikassa.

Työssä käytiin läpi keskeisimpiä eroja relaatiotietokantojen ja NoSQL-tietokantojen välillä erityisesti big data -analytiikan näkökulmasta. Näiden tietokantaratkaisujen keskeisimpiä eroja käsiteltiin työn tuotoksena syntyneessä taulukossa, joka on esitelty aikaisemmin kappaleen 3 yhteenvedon yhteydessä. Ensimmäinen tutkimuskysymys oli

1. Miten relaatiotietokannat ja NoSQL-tietokannat soveltuvat big data -analytiikkaan ja mitä eroja niillä on?

Tietokantaa valitessa vaihtoehtojen määrän suuruus voi tehdä tehtävästä hankalan, erityisesti jos tietämystä tietokantojen välisistä eroavaisuuksista ei entuudestaan ole. Jopa tieteellinen kirjallisuus on joissain tapauksissa keskenään eriävää mieltä siitä, kuinka hyvin tietokannat soveltuvat eri tarpeisiin. Työssä todettiin, että big data -analytiikan näkökulmasta uudet NoSQL-tietokannat tarjoavat enemmän optimoituja vaihtoehtoja erilaisiin käyttötarkoituksiin, parempaa tukea rakenteettomalle datalle sekä helpompaa tietokannan horisontaalista skaalautuvuutta. Nämä tekijät tekevät NoSQL-tietokannoista useimmissa tilanteissa ylivertaisia ratkaisuja big data -analytiikassa. Lisäksi NoSQL-tietokannat ovat pääsääntöisesti kustannuksiltaan edullisempia kuin relaatiotietokannat.

Työssä kuitenkin todettiin myös, että NoSQL-tietokantoihin liittyy useita heikkouksia, kuten toimittajien vähyys markkinoilla. Tämä voi tehdä NoSQL-tietokantojen hankinnasta ja tuen saamisesta niiden käyttöön ongelmatilanteissa haasteellista. Myös se, että NoSQL-tietokannat ovat hyvin uusia ja paljolti vasta kehitysvaiheessa, voi olla riski niiden luotettavuuden kannalta.

Tämän lisäksi NoSQL-tietokannoissa on relaatiotietokannoista poiketen heikko tuki transaktioille ja löysemmät vaatimukset tietokannan eheyden suhteen.

Työn perusteella voidaan todeta, että sopivaa tietokantaa valittaessa on tärkeä selvittää, minkä tyyppistä dataa tietokannan tulisi pystyä käsittelemään. Yhtä tärkeää on myös tunnistaa ne loogiset ja fyysiset toimenpiteet, joita tietokannan avulla käsiteltävälle datalle tulee pystyä toteuttamaan. Näiden tietojen avulla voi rajata, vastaavatko relaatiotietokantojen vai NoSQL-tietokantojen ominaisuudet paremmin tarvetta. Erityisesti NoSQL-tietokantojen eroavaisuudet toisistaan ovat todella merkittäviä, joten rajaaminen pelkästään tietokannan tyyppiin ei vielä riitä määrittelemään sopivaa vaihtoehtoa tarpeeksi tarkasti. Digitalisaation edetessä yrityksissä yhä pidemmälle onkin luonnollista, että yksi ainoa tietokanta ei pysty tyydyttämään kaikkia yrityksen tietovarastoinnin tarpeita, vaan useamman eri tietokannan samanaikainen hyödyntäminen on välttämätöntä, jotta mahdollisimman paljon erilaista tietoa pystytään hyödyntämään. Erilaisten tietokantojen yhteensopivuus vaatii ammattilaisen kokemusta ja osaamista, sillä valmiita kombinaatioita ei yksiselitteisesti voida määritellä. Lisäksi sekä uusien relaatio- että NoSQL-tietokantojen kehitys tulevaisuudessa pyrkii mahdollistamaan entistä paremman tuen horisontaaliselle skaalattavuudelle big datan käsittelyä varten, mikä tuo molempia tietokantatyyppejä lähemmäksi toisiaan.

Datan varastointiratkaisun hankinnan osalta työssä käytiin läpi pilvipalveluita sekä niihin liittyviä teknisiä ratkaisuja ja liiketoimintamalleja. Työssä käsiteltiin eri pilvipalveluratkaisuiden keskinäisiä eroja ja erityisesti niitä tekijöitä, jotka vaikuttavat sopivan ratkaisun hankintaan eri yritysten kohdalla. Työn toinen tutkimuskysymys oli

2. Mitä erilaisia ratkaisuja yrityksillä on datan varastointiratkaisujen hankintaan data-analytiikassa?

Työssä tunnistettiin erilaisiksi datan varastointiratkaisun hankintakeinoiksi yksityinen pilvi, julkinen pilvi sekä hybridipilvi. Työssä huomattiin, että sopivimman pilvimallin valinta tietovarastoinnin toteuttamiseksi analytiikan tarpeisiin riippuu vähintään yhtä paljon yrityksen ominaispiirteistä kuin itse analytiikasta. Julkisen pilven mahdollistama rajaton laskentakapasiteetti

ilman etukäteisinvestointeja on houkutteleva vaihtoehto erityisesti pienille yrityksille, sekä yrityksille joilla kokonaissuoritustehon tarve vaihtelee paljon. Suuria yrityksiä, joille tietoturva on tärkein prioriteetti, palvelee yksityinen pilvi, joka tuo virtualisoinnin tuomat edut käyttöasteen ja suorituskyvyn maksimoimiseksi, mutta säilyttää datan tarkasti varjeltuna yrityksen omissa tiloissa. Eriasteiset hybridiratkaisut mahdollistavat parhaiden puolien hyödyntämisen molemmista vaihtoehdoista. Kappaleen 4 yhteenvedossa esitelty nelikenttämatriisi kiteyttää työssä löytyneet vastaukset 2. tutkimuskysymykseen. Valinnanvara SaaS-, PaaS- ja IaaS -ratkaisuista mahdollistaa sopeuttamisen tietoteknisen infrastruktuurin ylläpitoon tarvittavan työpanoksen määrän ja lopullisen järjestelmäkokonaisuuden yksilöinnin välillä.

Työssä ei perehdytty kovin syvällisesti erilaisten NoSQL-tietokantojen teknisiin ominaisuuksiin ja niiden keskinäisiin eroihin big data -analytiikan kannalta. Myöskään Hadoopia ei työssä käsitelty tarkasti. Työhön liittyen voisi tehdä jatkotutkimusta Hadoopin käyttöönotosta big data-analytiikkaan, sillä Hadoop on hyvin laaja viitekehys erilaisille etenkin big dataan soveltuville datan varastointiratkaisuille. Lisäksi jatkotutkimuksena olisi luonteva tehdä vertailua NoSQL-tietokantojen välillä niiden soveltuvuusalueista big datan varastoinnin ja analytiikan tarpeisiin.

LÄHTEET

Assunção, M., Calheiros, R., Bianchi, S., Netto, M. & Buyya, R. 2015. Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*. Vol 79. s. 3-15.

Beach, C. 2014. Unstructured Data. *Journal of Accountancy*. Vol 217, nro 1. s. 46-51.

Biswas, S., Roy, A. & Chowdhury, M. 2013. Cloud Service: Cost Effective Strategy to Manage Cloud Services. *International Journal of Computer Applications.* Vol 71, nro 9. s. 33-38.

Boukhelef, D., Boukhobza, J. & Boukhalfa, K. 2016. A Cost Model for DBaaS Storage. *International Conference on Database and Expert Systems Applications*, s. 223-239.

Casado, R. & Younas, M. 2015. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and Experience*. Vol 27, nro 8. s. 2078-2091.

Chen, M., Mao, S., Zhang, Y. & Leung, V. 2014. Big data: related technologies, challenges and future prospects. Heidelberg: Springer. 89s.

Foster, E. & Godbole, S. 2016. Database systems: a pragmatic approach. Apress. 508s.

Gupta, R., Gupta, H. & Mohania, M. 2012. Cloud computing and big data analytics: what is new from databases perspective?. *International Conference on Big Data Analytics*. s. 42-61.

Grimes, S. 2008. Unstructured Data and the 80 Percent Rule. Breakthrough Analysis. [verkkoartikkeli] [Viitattu: 19.4.2017] Saatavissa: https://breakthroughanalysis.com/2008/08/01/unstructured-data-and-the-80-percent-rule/

Grolinger, K., Higashino, W., Tiwari, A. & Capretz, M. 2013. Data management in cloud environments: NoSQL and NewSQL data stores. *Journal of Cloud Computing: Advances, Systems and Applications*. Vol 2, nro 1.

Heino, P. 2010. Pilvipalvelut. Helsinki: Talentum Media. 267s.

Januzaj, Y., Ajdari, J. & Selimi, B. 2015. DBMS as a Cloud service: Advantages and Disadvantages. *Procedia-Social and Behavioral Sciences.* Vol 195. s. 1851-1859.

Levin, S. 2015. Choosing a Database for Analytics. Segment. [verkkoartikkeli] [Viitattu: 26.3.2017] Saatavissa: https://segment.com/blog/choosing-a-database-for-analytics/

Loshin, D. 2013. Big Data Analytics. From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph. Elsevier. 120s.

Moniruzzaman, A. & Hossain, S. 2013. NoSQL Database: New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison. *International Journal of Database Theory and Application.* Vol. 6, No. 4.

Romanek, E. 2015. Driving Analytics SaaS, PaaS, and IaaS with Managed Services: The Difference that Experts Make. Ironside Group. [verkkoartikkeli] [Viitattu: 27.3.2017] Saatavissa: https://www.ironsidegroup.com/2015/06/03/driving-analytics-saas-paas-and-iaas-with-managed-services-the-difference-that-experts-make/

Runkler, T. 2012. Data Analytics. Heidelberg: Springer. 126s.

Salo, I. 2010. Cloud computing-palvelut verkossa. Jyväskylä: WSOYpro. 168s.

Uckelmann, D., Harrison M. & Michahelles, F. 2011. Architecting the Internet of Things. Heidelberg: Springer. 347s.

Vaas, L. 2016. To SQL or NoSQL? That's the database question. Ars Technica. [verkkoartikkeli] [Viitattu 21.3.2017] Saatavissa: https://arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/

Zhao, L., Sakr, S., Liu, A. & Bouguettaya, A. 2014. Cloud data management. Heidelberg: Springer. 189s.