Lauri Nuutti Anton Rantanen

Big Datan louhinta ohjelmistokehityksessä

Tietotekniikan kandidaatintutkielma

13. kesäkuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Lauri Nuutti Anton Rantanen

Yhteystiedot: nuutti.rantanen@luukku.com

Työn nimi: Big Datan louhinta ohjelmistokehityksessä

Title in English: Big Data mining in software develoment

Työ: Kandidaatintutkielma

Sivumäärä: 20+0

Tiivistelmä: Nykypäivänä maailmassa kertyy dataa valtavia määriä. Big Data on tällä hetkellä ehkäpä yksi kunnianhimoisimmista datateknologioista. Tätä datamäärää voidaan tutkia ja valjastaa tieteellisiin ja kaupallisiin tavoitteisiin. Valtava data ei itsessään ole Big Dataa, mutta Big Data on valtavaa. Tavoitteena on kartoittaa Big Datan säilöntään, tietoturvaan, louhintaan ja saavuttamiseen liittyviin ongelmiin. Big Data on nykyaikaa ja on tullut jäädäkseen meidän iloksemme. Tutkimuksessa on käytetty päälähteenä aiheeseen liittyvää kirjallisuutta ja tulokset painottuvat kirjallisuudessa esiintyneisiin malleihin.

Avainsanat: Big Data, Big Datan kehitys, optimointi, Big Datan turvallisuus, Big Datan säilöntä

Abstract: Big Data is something that comes up once in a while. Like most data related topics Big Data is the most ambitious technology. We can use Big Data to help us record and process economic or scientific projects better than like ten years ago. Big Data is huge, big but not all huge amount of data is Big Data. Goals are to map uses for Storage for Big Data, Big Data security, Big Data mining and reaching Big Data. Big Data is modern thing and it's here to stay. Main sources are other literatur and results are based on that literatur.

Keywords: Big Data, Develoment of Big Data, optimation, Security of Big Data, Storage of Big Data

Taulukot

Sisältö

1 Johdanto

Big Data on nykyään kaikkialla ja Big Datalla voi tutkia populaatiota, ihmisten kuluttaja tottumuksia, ihmisten suhteita ja biologian ilmiöitä. Vaikkapa suuren määrän ihmisistä kauppaostoshistoria. Big Datan määritelmään kuuluu sen jatkuva kasvu eli ihmisen kauppahistoria muuttuu kun esimerkiksi henkilö ostaa jotain ja saadaan uusi alkio taulukkoon, mutta se kasvaa myös silloinkin kun henkilö asioi uudessa kaupassa ja haluataan laittaa Big Data aineistoon uusi sarake ”kaupat”. Eli Big Data kasvaa X ja Y akselien suhteen jatkuvasti.

Tutkielman päämääränä on vastata tutkimuskysymykseen, joka oli ”miten Big Dataa voidaan käyttää osana kasvavaa yhteiskuntaa ohjelmistokehittäjän näkökulmasta". Vaikka Big Data on kasvava osa tietoteknistä kehitystä, niin mitä ohjelmistokehittäjän kannattaa ottaa huomioon Big Datan hallinnasta.

Tässä kirjallisuuskatsauksessa esitellään Big Datan määritelmä lyhyesti, tutkitaan Big Datan ohjelmelmoinnin ongelmia ja esitellään ratkaisuja Big Datan ohjelmointiin käyttäen kirjallisuudessa olevia ratkaisuja. Katsauksen alussa keskitytään määrittelemään Big Data ja Big Datan louhinta. Kolmannessa luvussa on viittauksia ongelmiin ohjelmistokehittäjän suunnalta ja neljännessä luvussa esitetään ratkaisuja ja potenttiaalisia ratkaisumalleja. Lopuksi katsauksessa on yhteenveto, jossa esitellään tiivistetysti tutkielman hyötyjä ja käyttötarkoituksia.

2 Mitä on Big Data ja sen louhinta?

Mitä siis tarkoitetaan Big Datalla ja sen louhinnalla. Big Data on yleisesti sanottuna iso määrä dataa, tietoa, informaatiota, joka usein järjestelmällistä ja hyvin monimutkaista. Big Data kasvaa aivan koko ajan loputtomasti, joten sen säilönnässä on ongelma. Big Data orientoiduissa alustoissa on toinen ongelma, koska Big Datan monimutkaisuus, valtavuus, samankaltaisuus ja vaikeasti ennustettava. Tätä aukaistaan myöhemmin tutkielmassa.

2.1 Big Datan lyhyt historia

Big Data voidaan jäljittää 1990-luvun alkuun, mutta nykyään Big Data kiinnostaa varsinkin tietotekniikan, tilastotieteen, matemaattisen ja kauppatieteellisen alojen tutkijoita ja yrityksiä. Vuonna 1984 on alettu tutkimaan analogisen ja digitaalisen tallennustilan kokonaismäärää maailmassa ja tätä tutkimusta jatkettiin vuoteen 2007 asti. (Hilbert M & López, P. 2011) Tutkimus alkoi 1986 luvuista 2,6 eksabittiä (10^{18}) yhteensä tallennustilaa käytössä ihmiskunnalla vuonna 1986. Vuonna 1993 oli 15,8, vuonna 2000 oli 54,5 ja vuonna 2007 oli 295 eksabittiä. Ennen vuotta 2000 suurin osa datasta oli analogisilla tallennusvälineillä, kuten vhs-kasetit ja vinyylilevyt, mutta jo vuoteen 2007 tultaessa digitaaliset tallennusvälineet ovat vallanneet alaa merkittävästi analogisilta tallennusvälineiltä. (Hilbert, M. 2011) Tämä tutkimus on edellä auttanut käsitystä maailman tallennusmäärän kasvusta, joka on osaltaan johtanut Big Datan kehittymiseen.

2.2 Big Data

Big Datalla tarkoitetaan suurta määrää järjestettyä, vaihtelevaa tietoa. Big Dataa kertyy monesta eri lähteestä, kuten käyttäjien toiminnasta tietyllä verkkosivulla klikkaukset, sosiaalisessa mediassa kerätyt mallit ja erilaisissa sensoriverkoissa. Big Datalla voidaan tallentaa tietoa melkeinpä mistä vaan. (George, G. & Pentland , A. 2016)

Asiakas_id	*Sukunimi*	*ViimeisinkyntiSalessa*	*ViimesinkyntiPrismassa*	...
1	*Korhonen*	3.4.	25.3.	
2	*Virtanen*	2.4.	29.2.	
3	*Nieminen*	3.4.		
...				
n	*Jokinen*	14.1.	29.2.	

Taulukko 1. Ensimmäinen Big Dataa havainnollistama taulukko

Big Dataa ei pidä sekoittaa vain pelkkään suureen data määrään vaan sen erittelee suurista aineistoista se, että se menee perinteisten aineiston käsittelytapojen yli. Joten Big Datalle tarvitsee kehittää omia tapoja käsitellä Big Dataa.

2.3 Big Datan havainnollistaminen

Big Dataa voidaan havainnollistaa kuvin ja kaavioin. Seuraavan havainnollistetaan Big Datan valtavuutta, nopeutta ja määrästä taulukkojen avulla. Taulukossa on havainnollistettu yksinkertaisesti mahdolliset kasvusuunnat Big Datalle.

Taulukosta huomataan, että jokaisella rivillä on yksi mahdollinen asiakas suomalaisessa päivittäistavarakauppaympäristössä, mutta tämä ei olisi suoraan Big Dataa ellei rivillä olevalle asiakkaalle voida lisätä uusia sarakkeita kuvaamaan hänen käyttäytymistään. Vaikkapa taulukkoon voidaan lisätä uusia kauppavierailuja, asiakaan ostotottumuksia tai vierailuaikoja. Teknisestä näkökulmasta voidaan yhden solun sisältö määrittää miksikä tahansa tietotyypiksi. Esimerkiksi kokonaisluvuksi, merkkijonoksi, taulukoksi tai peräti tietokannaksi. Taulukkoon voidaan lisätä uusia sarakkeita jatkuvasti, koska Big Datan keräyksen nopeus vaatii lisäsarakkeita. Esimerkiksi, jos halutaan kerätä Big Dataan tiedon ”onko asiakas ostanut tarjouksesta tuotteen?” niin sarake lisätään Big Dataan.

Huomattavaa on kuitenkin, ettei tämä taulukko todellisuudessa ole näin pieni ja

Asiakas_id	*Sukunimi*	...	*Ostanut tarjoustuoteen*	...
1	*Korhonen*		*true.*	
2	*Virtanen*		*false.*	
3	*Nieminen*			
...				
n	*Jokinen*		*false.*	

Taulukko 2. Toinen Big Dataa havainnollistama taulukko

yksinkertainen vaan äärimmäisen monimutkainen. Asiakkaita voi olla miljardeja ja mahdollisia kauppoja tuhansia. Myös jokaisella rivillä voi olla paljonkin täysin tyhjiä soluja, koska joltain asiakkaalta puuttuu täysin data tietyltä kohtaa.

2.4 Big Datan louhinta

Tällä hetkellä on hieman ongelmallista lähteä tarkastelemaan monien eri palveluiden päivittäin saamaan ja prosessoimaa datan määrää. Nykyisillään esimerkiksi Google prosessoi monia satoja petatavuja dataa, Facebook luo yli 10 petatavua kuukaudessa lokeja ja Alibaba luo kymmeniä teratavuja päivässä vain netissä käytävistä ostoksista. (Chen Mao ja Liu 2014) Haasteena Big Datan louhinnalle on datan saavuttaminen ja prosessointi, datan yksilöllisyys, Big Datan säilytys ja Big Datan kaivuu algoritmit. Kehittämällä Big Datan louhintaa saadaan paremmin ja helpommin käyttöön Big Datan tarjoamat hyödyt kaupallisiin, tieteellisiin ja näiden alakategorioihin kuuluviin osa-alueisiin, kuten myyntiin, sosiaalisiin tutkimuksiin, biotieteisiin ja korkean energian fysiikkaan. (Talia Domeniko 2013)

2.5 Big Datan perusarkkitehtuurit

Kun teknologiamme kehittyvät eteenpäin kohti suurempia säilöntä metodeja ja tehokkaampia laskentatyökaluja niin Big Datan kerääminen, laskeminen ja säilöntä helpottuvat. Kerättävän ja saatavilla olevien datojen määrät kasvavat jatkuvasti ja täten Big Datan perusarkkitehtuureita tarvitsee kehittää. (Pop, Kołodziej & Di Mar-

tino 2016)

Tulevaisuudessa mahdollisesti Big Datan määrän on arvioitu tuplaantuvan vuosittain ja Big Datan käsittelyjärjestelmät voivat ratkaista Big Datan ongelmien lisäksi myös muidenkin tietoteknisten alueiden ratkaisuissa, koska Big Datan eri arkkitehtuurit voivat jo saavuttaa lähes rajoittamattomat laskenta- ja säilytysmahdollisuudet. (Pop Florin ym. 2016)

3 Big Datan louhinta ohjelmistokehittälle

Ensimmäinen ongelma Big Datan siirtämisessä ohjelmistokehityksen puolelle on sen massiivinen koko. Kuten määrittelyssä ongelmia ovat oikeiden algoritmien etsiminen Big Datan louhinnalle. Seuraavaksi esitellään tarkemmin Big Datan louhinnalle tyypillisiä ongelmia.

3.1 Big Datan saavuttaminen ja prosessointi

Normaalin datan saavuttamiseen tarvitaan neljä askelta esikäsittely, datan muuntaminen, datan louhinta ja mallien tulkitseminen. Esikäsittely on datan hakemista, siivoamista ja integraatiota. Kun data on esikäsitelty niin data muunnetaan helposti käsiteltävään muotoon. Datan muuntamisen jälkeen datasta louhitaan tarvittavat kuvien tulkkaukset, kuten mahdolliset mallit. Kun data on louhittu niin malleista voidaan tehdä hyödyllisiä kaupallisesti tai tieteellisesti. (Xu, Jiang, Wang, Yuan & Ren 2014)

Ensimmäinen ongelma, joka tulee vastaan on kuinka Big Dataa kerätään ja toiseksi miten sitä prosessoidaan. Ohjelmistokehittäjälle voi koitua ongelmaksi vaikka miten mahduttaa hyvin paljon dataa käytössä oleviin muisteihin. Koska fyysiset muistit ovat suhteessa pieniä verrattuna useisiin Big Datan tietokantoihin niin voidaanko hajauttaa tietomassa järjevästi. Tapana voidaan hajauttaa tietokanta useampaan osaan ja tallentaa ne eri puolille. (Xindong, Xingquan Z, Gong-Qing W, & Wei D 2014)

Big Data on yhdellä tavalla saavutettavissa laskemalla Big Dataa. Yleisesti Big Dataa ei voida valtavuuden takia laskea käsin, koska Big Data laajenee jatkuvasti moneen suuntaan. Vaan Big Dataa joutuu laskemaan automatisoidusti vaikkapa koneellisesti. Tutkittavia asioita Big Datan koneellisessa laskennassa ovat laskentajärjestelmien hinta eli miten voidaan järkevästi tasapainottaa sijoitukset, tuotot ja järjestelmän pitkäikäisyys. (Pop Florin ym. 2016)

3.2 Big Datan säilöntä

Louhittu Big Data tarvitsee säilön, missä sitä voidaan säilyttää, mutta Big Datan valtavuuden vuoksi säilöntä on haaste. Tämä haaste on moninainen, johon kuuluvat matalan tason tiedostojärjestelmiä ja loogisia tietokantoja ja niiden ylläpitoa. (Pop Florin ym. 2016)
Säilytystilan tarvitsee olla mahdollisimman kestävää, johon tämän aikakauden suurimmat datan sijoituspaikat perustuvat magneettisiin säilöntätekniikoihin, kuten kovalevy ja vielä jossain määrin magneettinauhat. Nykyään ongelmana on teknologian jatkuva kehitys, suuri energian kulutus ja kovalevyjen, Hard disk drives, rajallisen käyttöiän takia. (Gu, Li, & Cao 2016)

On tärkeää ymmärtää, että nämä edellä mainitut seikat, joista varsinkin nykyisten tallennusratkaisujen rajallinen käyttöikä vaikeuttaa Big Datan säilöntää merkittävästi ja se pitää ottaa huomioon Big Datan louhinnassa.

3.3 Big Datan tietoturva

Tämä osa ei koske niin läheisesti ohjelmistokehittäjän näkemystä Big Dataan, mutta on silti tärkeä asia käsitellä. Big Datan yksilöllisyydellä tarkoitetaan yksityisen henkilön tietoturvaa. Voidaanko yksityisen ihmisen tietoja käyttää Big Datan omistajan mielen mukaan. Kuka voi omistaa ihmisten tietoja?

Tässä on huomioitava ohjelmistokehittäjänä, että mitä dataa menee kenellekin ja mihin tarkoitukseen. Toinen asia, jolla voidaan varmistaa tietoturvallinen Big Data, on tieto siitä millä domainilla ja ohjelmalla ollaan. (Xindong W ym. 2014)

Big Datajärjestelmät voivat sisältää tietoja mm. yritysten tiedoista, sairaanhoitotiedoista, mainonnasta ja valtioiden tilastoista. Big Datan tarvitsee olla tietoturvallista, että kyseiset seikat pysyvät vain niiden oikeuksien haltioilla. Vaikka paperilla olisi helppoa pitää Big Data tietoturvallisena niin sen valtavuus, nopeus ja sen käsitellyn ja keräyksen nopeus vaikeuttavat Big Datan kriitisen datan pitoa salassa. (Pop Florin ym. 2016)

Onko olemassa konkreettisia keinoja Big Datan pitämiselle tietoturvallisena? Vastausta voidaan etsiä ongelmaan kryptografian, pilvipalvelujen, tietoverkkojen, kompleksien järjestelmien ja simulaatioiden kautta. (Pop Florin ym. 2016)

Big Datan yksityisyyden suoja on uhattuna ja suojaavat teknologiat ovat vielä lasten kengissä. Dataa kertyy erittäin nopeasti, mutta samaan aikaan datan yksityisyyttä ei tarkastella tarpeeksi hyvin turvallisuuden kannalta. Turvallisuudesta puuttuvat hyvä valvonta, hyvä tekninen tuki ja hyvät valvontatyökalut ja datalla on heikkous tietohävikille. (Dongpo Zhang 2018)

3.4 Big Datan louhinta-algoritmit

Big Datalle on kehitetty ja ollaan kehittämässä tehokaampia ja nopeampia kaivuu algoritmejä. Yksi vaihtoehto on hajautettujen tietokantojen käsittely algoritmi, jolla voidaan käsitellä Big Dataa, joka on hajautettu useampaan tietokantaan.

Ongelmia mihin voi törmätä ohjelmistojen kehittämisessä ovat väljän, epävarman ja epätäydellisen datan käsittelyalgoritmien käsittelemisessä. (Xindong W ym. 2014) Epävarma dataa syntyy, kun datakenttä ei enää olekaan deterministä vaan kentässä alkaa syntyä erinäisiä virheitä. Epävarmaa dataa alkaa yleensä kertymään, kun työkalut itsessään muodostaa virhettä dataan. Esimerkiksi GPS-välineet ovat usein epävarmoja, koska niille ei ole olemassa kunnollista ja varmaa teknologista ratkaisua. Jotkin käyttäjät altistavat laitteen virheille, ettei heidän yksityistietonsa päädy täysin oikeellisena Big Dataan, koska heillä on huolta tietokannan tietoturvasta. (Xindong W ym. 2014)

Epätäydellinen data luo haasteita taas toisaalla. Epätäydellistä dataa syntyy kun vaikka sensori hajoaa ja se ei tuota mitään dataa tai järjestelmän ominaisuus estää datan saannin tietyllä hetkellä. Nykyään Big Datan kehittäjät osaavat ottaa huomioon mahdolliset Big Datan puuttuvat tiedot ja jotkin estimoivat yleisesti saadusta datasta tai koneoppialgoritmeilla, mitä puuttuva data voisi olla. (Xindong W ym.

2014)

Nykypäivän kompleksidata ajaa Big Dataa eteenpäin jatkuvasti, koska internetistä kertyy jatkuvasti enemmän ja nopeammin dataa. Esimerkiksi erinäiset dokumentit, sosiaaliset verkostot, viestintäverkot, ja tietoliikenne lasketaan kompleksiseksi dataksi. Kompleksidata lisää haasteita louhia Big Dataa nopeammin, mutta tutkijat ovat kehittäneet algoritmin, millä saadaan sosiaalisesta verkosta, kuten Twitteristä ajantasaista tietoa maailmalla tapahtuneista tapahtumista hyvinkin tarkasti. Kompleksisen datan käyttö on vielä haaste Big Datan louhijoille, koska jos meillä on kaksi toimijaa, jotka omistavat kompleksista dataa ja haluavat louhia toisen dataa, niin se nelinkertaistaan tarpeen hallita Big Datan valtavuutta eli säilöntätilaa ja louhintatehoa. (Xindong W ym. 2014)

Big Dataa kertyy kaivuun aikana valtavasti epäjärjestäytynyttä dataa, joka tarvitsee tehokkaita algoritmeja, että dataa voidaan käyttää hyödyksi vaikkapa kaupallisissa tarkoituksissa. Elämme nopean tiedonkulun aikaa ja Big Datan tarvitsee olla nopeampaa, että saisimme tuloksia nopeammin käyttöön. (Talia Domeniko 2013)

Yhtenä keinona on jakaa dataa pienempiin klustereihin ja käsitellä klustereita erikseen, tätä kutsutaan hajonta-algoritmiksi. Hajonta-algoritmeilla pyritään hajauttamaan Big Data pienempiin osiin, joihin kuitenkin on pyrkimys jättää tarpeeksi dataa Big Datasta, johon on säilytetty Big Datan esittämä informatiikka. (Fahad, Alshatri, Tari, Alamri, Khalil, Zomaya, Foufou & Bouras 2014)

4 Ratkaisumalleja

Yleisesti Big Datan louhinnan mahdollisista ratkaisumalleista. Big Datan louhinnassa ei aika ole kriittinen tekijä, kuitenkin pitää tutkia ja mahdollisesti optimoida kaivuun aika. Big Datan luonteeseen kuuluu, että se on jatkuvasti kasvavaa ja täten tarvitsee enemmän ja enemmän laskentatehoa tietokoneilta. (Pop Florin ym. 2016)

Big Datalla on ongelmia, kuten säilönnän, valtavuuden, useiden algoritmien kehittämisen ja tietoturvan kanssa. Osa näistä sivuavat toisiaan samalla tavalla kuin venn-diagrammin eri osat ja toiset taas ovat enemmän yksin, mutta kuitenkin nämä esitetyt ongelmat kuuluvat teoreettisella tasolla samaan diagrammiin, missä asiat liittyvät toisiinsa ja samalla ovat myös itsenäisiä asioita.

Ratkaisumalleja-luvussa tarkistellaan ongelmia niiden asioiden pohjalta, miten tietyt ongelmat voidaan ratkaista taloudellisesti, tehokkaasti ja tietoturvallisesti ja nämä asiat hyödyttävät tiede- ja talousmaailmaa esittämällä ratkaisumalleja, joita voidaan käyttää tieteessä ja kaupallisesti. Big Datan ratkaisumalleja voidaan käyttää myös muiden valtavien tai muiden Big Datan kaltaisten datajoukkojen ratkaisussa.

4.1 Big Datan saavuttaminen

Big Datan suurin ongelma on sen valtava saatavuus ja siihen onko olemassa keinoja millä saavutetaan Big Dataa. Yksi tällainen keino on sopivan louhinta-algoritmin, joka perustuu yleishyödylliseen rinnakkaisen ohjelmointimetodin löytämiseen.(Xindong W ym. 2014) Toinen laskentatehoa ja säilyttämistä rajoittava tekijä on infrastruktuurin hinta eli pitää olla taloudellinen tuki ylläpitää tarpeaksi laskentatehoa ja säilytystilaa. (Pop Florin ym. 2016) Ja samaiseen asiaan liittyen pitää olla varma, että laskentateho- ja säilytysarkkitehtuurit tukevat Big Datan valtavaa saatavuutta. Monet asiat näissä voivat vaikeuttaa Big Datan saatavuutta, kuten algoritmien laatu, tiedonsaannin oikeus ja käsittelyteknologiat. (Pop Florin ym. 2016)

Big Datan saavuttamiseen löytyy myös haasteita sen säilytykseen liittyen, koska mi-

tä nopeammin, epäjärjestelmällisemmin ja valtavammin Big Dataa kertyy niin, että Big Data tarvitsee enemmän ja enemmän säilytystilaa. (Talia Domeniko 2013)

4.2 Big Datan säilytys

Big Dataa voidaan säilyttää monella eri käytännön ja teorian tasolla. Big Dataa on hankala säilyttää vain yhdessä paikassa sen valtavuuden vuoksi. Yksi hyvä ratkaisu Big Datan säilytykseen on pilvipalvelujen käyttö. Pilvipalvelujen avulla voidaan jakaa helposti säilytystilaa useampaan paikkaan ja myöskin pilvipalvelut auttavat Big Datan analyysissa ja laskennassa. (Talia Domeniko 2013)

Tänä päivänä suurin osa Big Datasta on säilöttynä erilaisten pilvipalvejuden avulla. Suurimmat yrityksen, jotka tarjoavat Big Datan pilviratkaisuja ovat mm. Google, IBM ja MigML. On oletettavaa, että Big datan säilytys ja analysointi jatkuu ja kehittyy pilvipohjaisena. (Talia Domeniko 2013)

Big Datan säilönnän loogisella tasolla tietokannat eivät skaalaudu tarpeeksi nopeasti ja tehokkaasti, että niitä voitaisiin käyttää järkevästi Big Datan säilöntään. Nykyään suurin tietokantarakenne perustuu NoSQL-tietokantoihin, jota optimoimalla poistamalla raskaita ominaisuuksia, kuten eheys, tiedustelujen optimointi, lukitus ja kaupallinen tuki. (Pop Florin ym. 2016)

4.3 Big Datan valtavuuden hallinta

Valtavuuden hallinnalla tarkoitetaan tässä yhteydessä Big Datan säilöntää ja laskentaa. Big Datan käyttämä tila on erittäin valtava ja tämän lisäksi Big Dataa kertyy jatkuvasti lisää. Yksi mahdollinen ratkaisu on jakaa data moneen eri paikkaan ja käyttää pilvepalveluita tai muuta vastaavaa säilönnässä ja laskennassa. (Talia Domeniko 2013)

Nykyään on mahdillista luoda tarvittavat pilvipalvelut Big Datan säilöntään ja laskentaan. Yleinen järjestelmä perustuu niin sanottuun palvelusuuntautuneeseen arkkitehtuuriin, jonka perusperiaatteita ovat joustavan, modulaarisen ja yhteen toimi-

van ohjelman rakentamiseen. (Pop Florin ym. 2016) Tämän arkkitehtuurin voi esittää erillisenä web-palveluna. Web-palvelut ovat funktioiden kokoelma, jotka ovat pakattu yhdeksi kokonaisuudeksi ja julkaistu tietoverkossa ja noudattavat jotain yleistä standardia. Web-palvelut voivat suorittaa itsenäisesti laskutoimituksia ja palvelujen komponentit tietävät vain, että toiset komponentit palauttavat odotetun tuloksen. Tästä alkaa muodostua pilvipalvelu, jossa on useita eri palveluja yhdessä ja saadaan moninkertaistettua laskenta- ja säilöntäkapasiteetti Big Datalle. (Pop Florin ym. 2016)

Pilvipalvelut mahdollistavat Big Datalle ja muille massiivisille laskenta- ja säilöntäongelmille ratkaisuja, joita ei muilla teknologioilla voi järkevästi ja tehokkaasti ratkaista. (Pop Florin ym. 2016)

4.4 Big Datan tietoturva

Big Datan tietoturva on monimutkainen asia, koska Big Dataan mahdollisesti pääsevät käsiksi mm. datan omistajat, järjestelmän ostajat ja kolmannen osapuolen organisaatiot. Datan määrä ja nopeus lisäävät riskiä järjestelmän tietoturvalle. (Pop Florin ym. 2016)

Big Datan tietoturvalle tyypillisiä näkökulmia ovat käyttäjäturvallisuus, dataturvallisuus ja järjestelmäturvallisuus. Ongelmaksi muodostuu usein, ettei Big Datan dataa on hajautettu useampaan tietokantaan. Tämä on kannattavaa siinä suhteessa, että valtava määrä dataa tarvitsee tilaa. Tämä kuitenkin lisää riskiä Big Datan päätymisen ulkopuoliselle taholle, jolle data ei kuulu. (Pop Florin ym. 2016)

Big Datalla on neljä erilaista käyttäjärooleja, datan tarjoaja, datan kerääjä, datan louhija ja päätöksen tekijä. Tarjoaja päättää datan arkaluontoisuudesta ja sen jakelusta, kerääjä kerää datan arkaluontoista lähteistä ja siistii datasta arkaluontoiset tiedot pois, louhija käyttää louhinta-algoritmeja siistitystä datasta ja kapseloi datan julkisturvalliseksi. Lopuksi päätöksen tekijä tarkistaa datan ja päättää onko data valmis käytettäväksi. (Lei Xu ym. 2014)

Big Datalle on luotava mahdollisen turvallinen ympäristö, josta ulkopuoliset eivät pääse sisään. Yksi tapa on toteuttaa täydellinen valvonta sosiaalisten verkkojen datalle. Nykyään melkeinpä jokaiselta kehittyneen maan kansalaiselta löytyy ainakin jokin sosiaalisen median väline ja Big Datan louhijat voivat viedä tietoja käyttäjistä vahingossa tai tahallisesti. Tässä pitää olla täysi valvonta kenelle dataa annetaan, jotta sitä ei päädy vahingossa ulkopuoliselle. (Dongpo Zhang 2018)

Big Datan tietoturvaa voidaan parantaa myös luomalla yhteisiä sääntöjä ja lakeja. Nykypäivänä on säädetty lakeja, joiden tarkoitus on suojella kansalaisten yksityisyyttä. Esimerkiksi Kiinassa vaikka, jokin viranomainen tietää toisesta ihmisestä tietoja niin tietoja ei saa antaa eteenpäin. (Dongpo Zhang 2018) Jotkin maat ovat myös kehittäneet toimijoita, joiden tarkoitus on tarkkailla yksityisten henkilöiden yksityisyyttä ja heidän datansa turvallisuutta. Kuitenkaan vieläkään ei kansalaiset tiedä tarpeeksi hyvin mihin heidän yksityisiä tietoja käytetään ja kuka kerää, mutta asiat ovat menossa parempaan suuntaan, kun kansalaisia valistetaan datan olemassa olosta ja siihen liittyvistä turvallisuusriskeistä. (Dongpo Zhang 2018)

4.5 Big Datan louhinta-algoritmit

Big Datan analysointiin ja louhintaan voidaan luoda monenlaisia algoritmeja. Algoritmeja jaetaan monella tavalla eri kategorioihin ja eri luokkiin.

Algoritmit voidaan jakaa kolmeen luokkaan, staattiset, puolistaattiset ja dynaamiset. Staattiset algoritmit pysyvät jatkuvasti samoina toisin kun puolistaattiset, joita muokataan päivittäin tai viikoittain. Dynaamiset algoritmit yrittävät muokata algoritmeja jatkuvasti louhinnan aikana. (Pop Florin ym. 2016)

Hajonta algoritmit ovat myös nostamassa päätään Big Datan louhinnassa. Hajonta-algoritmit ovat oppivia algoritmeja, joiden perusperiaate on jakaa data pienempiin klustereihin, joissa on saman kaltaista dataa keskenään. Hajonta-algoritmeja voidaan jakaa erilaisiin luokkiin, kuten osittainen, hierarkkinen, tiheydellinen, ruudukollinen ja mallipohjainen. (Fahad A ym. 2014)

Näissä hajonta-algoritmeissa on hyviä, että huonoja puolia. Osittaispohjaisessa pyritään jakamaan osat, jotka ovat samalla klustereita ja jokainen osa sisältää vähintään yhden objektin ja jokainen objekti kuuluu vain yhteen osaan. Hierarkiallispohjaisissa data jaetaan hierarkiallisesti etäisyyden mukaan puurakenteeseen, jossa lehdet ovat yksittäistä dataa. Tiheyden mukaan jaettu data puolestaan jaetaan lähimpänä olevan naapurin viereen, jonka sisältämä data on mahdollisimman lähellä naapuriaan. Ruudukollisessa data on jaettu ruudukoiksi, jotta data voidaan ratkaista tilastollisesti yhdessä ruudukossa kerrallaan. Suurin ongelma ruudukollisessa on, jos ja kun Big Data voi sisältää paljon epämääräistä dataa, jota on vaikea lajitella ruudukkoon järkevästi niin ruudukko ei tuota halutta tulosta. (Fahad A ym. 2014)

Mallillinen Big Datan klusterointi on keino optimoida klusterointia perustamalla oletuksen, että Big Datasta löytyy pohjalta jokin todennäköisyysjakauma. Samalla se olettaa, että klusterien määrä vastaa normaaleja tilastomenetelmiä. Täten se luo kilpailukykyisiä klustereita muihin klusterointimetodeihin verrattuna. Mallillisia klusterointimetodeja voidaan käyttää tilastollisesti ja myös neuroverkkojen laskennassa. (Fahad A ym. 2014)

5 Yhteenveto

Big Data on sen alkuajoilta asti kiinnostanut tiedemiehiä ja yrityksiä. Ei pelkästään sen hankala tutkittavuus tai sen kaupallinen arvo vaan myös Big Datan tietoturva on ja tulee kiinnostamaan ihmisiä vielä pitkään. Tiedon tallennus on lähtenyt pienestä määrästä dataa, ja nykyään voidaan miettiä onko jokin data pientä, suurta vai peräti Big Dataa.

Usein ihmisiä kiinnostaa jonkin asian käyttömahdollisuudet ja Big Datan kohdalla käyttömahdollisuuksia on useita kuten kaupallinen, tieteellinen ja näiden alle sijoittuvat välimuodot. Big Dataa voidaan hyödyntää kauppojen kehittämiseen, suurien tilastojen tekemiseen, biotieteissä saadaan suuret määrät dataa ja valtioiden verojärjestelmissä. Tätä varten on tärkeää osata louhia, säilöä, tutkia ja käsitellä Big Dataa oikein ja johdonmukaisesti. Haasteita tässä kuitenkin on valtavuuden, nopeuden, epäjärjestelmällisyyden ja monimutkaisuuden käsittely.

Ohjelmistokehittää pitäisi Big Datan louhinnasta ymmärtää sen valtavuus, tietoturvallinen käsittely ja louhinnasta muodostuneet mahdollisuudet. Louhinta on kaikista asioista se tärkein, koska ilman tuloksia Big Datalla ei ole mitään kauppallista tai tieteellistä arvoa. Louhijan on kuitenkin ensin ymmärrettävä Big Datan valtavuuden hallinta ja ottaa huomioon tietoturvaan liittyvät lait ja määräykset.

Kirjallisuutta

Chen, M., Liu Y. & Mao, S. 2014. *Mobile Netw Appl*. Teoksessa Mobile Netw Appl. Springer US (toim.) Saatavilla WWW-muodossa 10.1007/s11036-013-0489-0

Fahad A., Alshatri N., Tari Z., Alamri A., Khalil I., Zomaya A.Y.,Foufou S.& Bouras A. 2014. *A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis*. 10.1109/TETC.2014.2330519 Julkaistu: IEEE Access .Viitattu 29.5.2019.

George, G. & Pentland. 2016. *Big Data and Management*. ACAD MANAGE J, 57, s. 321–326. 10.5465/amj.2014.4002

Gu, M, Li, X & Cao, Y. 2016. *Resource Management for Big Data Platforms*. 10.1038/lsa.2014.58 Light: Science and Applications.Viitattu 15.5.2019.

Hilbert,M & López, P. 2011 .The World's Technological Capacity to Store, Communicate, and Compute Information 10.1126/science.1200970

Talia, D. 2016. *Clouds for Scalable Big Data Analytics*. 10.1109/MC.2013.162 Julkaistu: Computer (Volume: 46 , Issue: 5 , May 2013).Viitattu 21.5.2019.

Pop F, Kołodziej J & Di Martino B, K. 2016. *Resource Management for Big Data Platforms*. Pääosin sivut 3-8, 55-57 & 241-246. 10.1007/978-3-319-44881-7 Springer International Publishing. Viitattu 15.5.2019.

Xindong W, Xingquan Z, Gong-Qing W, & Wei D. 2014. *Data Mining with Big Data* . IEEE Transactions on Knowledge and Data Engineering, 26, s. 97–107. 10.1109/TKDE.2013.109

L. Xu, C. Jian, J. Wang, J. Yuan & Y. Ren. 2014. *Information Security in Big Data: Privacy and Data Mining*. 10.1109/ACCESS.2014.2362522 Julkaistu: IEEE Access .Viitattu 21.5.2019.

Zhang, D. 2016. *Proceedings of the 8th International Conference on Management and Computer Science*. 10.2991/icmcs-18.2018.56 Julkaistu: Atlantis Press .Viitattu 30.5.2019.