Datan visualisoinnin kokemus virtuaalitodellisuudessa

Sakari Hassi

Tampereen yliopisto  
 Informaatiotieteiden yksikkö

Tietojenkäsittelytieteiden tutkinto-ohjelma

Pro gradu -tutkielma   
 Ohjaaja: Harri Siirtola  
 Syyskuu 2017

Tampereen yliopisto  
Informaatiotieteiden yksikkö

Tietojenkäsittelytieteiden tutkinto-ohjelma  
Sakari Hassi: Big Datan visualisointi virtuaalitodellisuudessa  
Pro gradu -tutkielma, 14 sivua, 4 liitesivua

Syyskuu 2017

Abstrakti tulisi kirjoittaa seuraavana, kun tutkimuskysymykset ja sisältö selkeentyneet.

Avainsanat ja -sanonnat: Big Data, virtuaalitodellisuus, visualisointi.

SISÄLLYSLUETTELO

SISÄLLYSLUETTELO 3

1. JOHDANTO 4

2. BIG DATA 5

2.1 Big Datan määrittely 5

2.2 Big Datan kuudes V – Visualisointi 7

2.3 Big datan hyödyt ja ongelmat 9

3. TIEDON VISUALISOIMINEN 12

3.1 Visualisointi yleisesti 12

3.2 Visualisoinnin tyypit 13

3.2.1 Parallel coordinates 15

3.2.2 Star coordinates 16

3.2.3 Tree map 18

3.3 Visualisoinnin työkalut 19

3.4 Visualisoinnin prosessi 20

3.5 MapReduce 21

4. VIRTUAALITODELLISUUDEN HYÖDYNTÄMINEN 23

4.1 Virtuaalitodellisuus aikaisemmin 23

4.2 Big Datan visualisointijärjestelmät 24

4.3 Yhteenveto aikaisempien järjestelmien pohjalta 27

4.4 HTC Vive 28

5. TESTIJÄRJESTELMÄN KEHITYS 30

6. TESTIJÄRJESTELMÄN KÄYTTÄJÄTESTAUS 32

7. YHTEENVETO 32

7.1 Käyttäjätestauksen tulokset 33

7.2 Loppuyhteenveteo 33

Kirjalliset viitteet 33

1. JOHDANTO

Tieto on aina ollut merkittävässä roolissa ja sillä on ollut strategista arvoa. Pinte [2006] näkee, että nykypäivän tärkein varanto ei ole enää materia vaan varallisuus perustuu aineettomaan omaisuuteen, jota ohjataan tiedolla. Myös Buyya ja kumppanit [2009] ovat sitä mieltä, että tietojenkäsittelyn kautta tuotettu tieto tulee olemaan viides hyödyke sähkön, veden, bensan ja teleliikenteen lisäksi. Big Data on ollut jo muutamia vuosia kuuma puheenaihe tutkimus- ja yritysmaailmassa. Nykyään onkin vaikea olla huomaamatta Big Datan mainintaa monissa asiayhteyksissä. Alati kasvava luodun tiedon määrä ja sensoriverkkojen kehitys on johtanut tilanteeseen, jossa yhä useammasta fyysisen maailman tapahtumasta jää digitaalinen jälki. Tätä tietoa on nyt yhä laajemmin alettu arvostamaan ja käyttämään osana päätöstentekoprosesseja. Big Datan ongelmana on kuitenkin sen käyttöön valjastaminen. Tiedon tarkasta sisällöstä ei voida olla varmoja ja tiedon sisäisten yhteyksien löytäminen sekä analysoiminen ovat työläitä prosesseja. Näiden tekijöiden lisäksi tiedon käsittely vaatii paljon säilytyskapasiteettiä sekä Big Datan käsittelyyn luodun ohjelmistoarkkitehtuurin. Big Datan käsittelyn ongelmat huomioiden, Brenton Faber totesikin vuoden 2013 Association of Teachers of Technical Writing -konferenssissa meidän teknologisten mahdollisuuksien kerätä tietoa ylittävän meidän analyyttisen kapasiteetin luoda uutta ja merkityksellistä tietoa tämän aineiston pohjalta [Faber, 2013].

Kasvavien data määrien käsittelyyn ja hyödyntämiseen on pyritty etsimään ratkaisua uusista teknologia-alan innovaatioista. Virtuaalitodellisuus ei ole tällä saralla uusi keksintö ja virtuaalitodellisuuteen pohjautuvia ratkaisuita on jo muutamia vuosia hyödynnetty erityisesti peleissä sekä markkinoinnillisena työkaluna erityisesti realististen kokemuksien ja aistimuksien välittämiseen käyttäjälle. Virtuaalitodellisuutta on myös esitetty ratkaisuna alati kasvavien ja kompleksisten datakokonaisuuksien esittämiseen. Dataa visualisoivia ja virtuaalitodellisuutta hyödyntäviä järjestelmiä ei kuitenkaan juuri toistaiseksi ole saatavilla. Tässä tutkimuksessa tarkoituksena onkin arvioida virtuaalitodellisuuden soveltuvuutta datan visualisointiin ja mitä haasteita tai mahdollisuuksia virtuaalitodellisuus tuo verrattuna normaaliin työpöytäympäristöön. Toisena tutkimusnäkökulmana on selvittää miten käyttäjät kokevat datan visualisoinnin kokemuksena verrattuna normaaliin visualisointitapoihin. Yhteenvetona pyritään avaamaan syitä sille miksi ennakkonäkemyksistä huolimatta virtuaalitodellisuuden ratkaisut eivät ainakaan toistaiseksi ole saavuttaneet suurta suosiota Big Datan visualisoimisessa.

*”Our technical ability to gather data exceeds*

*our analytical capacity to make meaning from this data.”*

*Brenton Faber, 2013*

2. BIG DATA

2.1 Big Datan määrittely

Big Data on käsitteenä yhä varsin uusi ja sen määrittelyssä ei ole vielä saavutettu täyttä konsensusta. Joidenkin mielestä käsite kuvaa vain suurta datajoukkoa, kun taas joidenkin mielestä käsite on huomattavasti moniulotteisempi. Määritysten kehittäminen on erityisesti keskittynyt kuvaamaan, mikä on tarpeeksi suurta, jotta se laskettaisiin Big Dataksi. Vuonna 2014 Berkeleyn School of Information [Dutcher, 2014] teki kyselytutkimuksen ja pyrki selvittämään eri alojen asiantuntijoiden vastausten pohjalta yhteisen määrityksen Big Datalle, mutta sai vastaukseksi 40 jokseenkin toisistaan poikkeavaa määritelmää. Yhteneväisyytenä näille määrittelyille pystyttiin kuitenkin toteamaan datan massiivisuuden, monimuotoisuuden ja sen tiedon muutoksen nopeuden. Nykyään Big Datan määrittelyssä poikkeuksetta esiintyykin ainakin kolmen V:n määritelmä: tiedon määrä (*Volume),* tiedon nopeus *(Velocity)* ja tiedon monimuotoisuus (*Variety).* Tämä voidaankin pitää pohjamäärittelynä Big Datan kuvaamiseen. Laney [2001] alusti Big Datan kolmen V:n määritelmän kuvatessaan yrityksen tiedonhallinnan kasvavia ongelmia toistaiseksi vielä kuitenkaan mainitsematta Big Datan käsitettä osana esitystään. Varsinaisesti Big Data käsite esiteltiin vasta vuonna 2005, O’Reilly median jäsenen, Roger Magoulasin toimesta, hänen pyrkiessä kuvaamaan suurta tiedon määrää. Nämä kolme tekijää on kuitenkin vahvasti kartoitettu aihetta käsittelevissä tutkimuksissa ja ne voidaan nähdä merkittävimpinä tekijöinä Big Datan käsittelyssä.

* *Volume,* Tiedon määrä on kasvanut niin suureksi, että käsiteltävä data ei enää mahdu analysoinnissa käytettävien tietokoneiden muistiin, joten Big Datan käsittely vaatii erityisesti sille luodun järjestelmäympäristön. Esimerkkinä International Data Corporation [2014] toteaa ennustuksessaan datan määrän tuplaantuvan kahden vuoden välein vuonna 2015 datan määrä ollessa 8 zettatavua (1021). Tämän kaavan pohjalta vuonna 2020 dataa on ennustettu olevan jo 44 zettatavua.
* *Velocity,* kuvaa tiedon muuttumisen ja liikkumisen nopeutta. Tieto on nykyään reaaliaikaista ja tiedon analysointi tulisi suorittaa nopeasti sen tuottamisen jälkeen, sillä muuten tieto menettää nopeasti merkityksensä. IBM:n teettämien arvioiden mukaan vuonna 2018 globaalisti internet-protokollaa hyödyntävien laitteiden kaista on 50 000 gigatavua sekunnissa.
* *Variety,* Tiedon sisältö on muuttunut pelkistä numeroista ja tekstistä rakenteettomaan suuntaan eri laitteiden tuottaessa rakenteellisesti erilaista dataa. Tällöin käsiteltävä tieto vaatii osakseen metatietoja, joiden avulla tiedon sisältö pystytään kuvaamaan. Tiedon sisällön tarkka ymmärtäminen onkin yksi avaintekijöistä Big Dataa hyödynnettäessä.



*Kuva 1. Big datan viisi V:tä*

*(*[*https://www.omnivex.com/company/blog/what-is-big-data*](https://www.omnivex.com/company/blog/what-is-big-data)*)*

Laneyn määrittelemien alkuperäisen kolmen V:n päälle on useissa tutkimuksissa [Hadi et al., 2014 ja Rajaraman, 2016] esitelty yhä kaksi jatkomäärittelevää V:tä lisää: *Veracity* ja *Value* (Kuva 1).

* *Veracity,* Kerätty data ei välttämättä ole todenmukaista. Data saattaa olla huonosti organisoitua tai puutteellista, jolloin sen sisältöön ei voida täysin luottaa ja tietoa ei voida käyttää osana luotettavaa päätöksentekoa. Erityisesti autonomisten, tietokonealgoritmien avulla ohjattujen, *Cyber-Physical-Systems (CPS*) järjestelmien kohdalla sensoridatan epäluotettavuus on arvioitu suurimmaksi verkon toimintaa estäväksi tekijäksi [Tolle et al., 2005, Szewczyk et al., 2004].
* *Value,* Tieto itsessään on arvotonta, ellei sitä pystytä prosessoinnin kautta hyödyntämään päätöksenteossa ja toiminnassa. Big Datan kohdalla ongelma on tiedon suuri määrä, mikä tulisi prosessoida ennen kuin siitä voidaan luoda arvoa. Perinteinen työasema ei itsessään riitä prosessoinnin suorittamiseen vaan käsittely vaatii hajautetun tiedonhallinta – ja prosessointiympäristön. Tällä hetkellä parhaimman alustan tähän tarjoaa *Hadoop* (*Hadoop Distributed File System)* sekä *MapReduce (Distributed computation framework)* [Shvachko et al., 2010].

Tässä tutkimuksessa Big Datan määritelmä nähdään kuitenkin Moninon [2016] antaman yleisemmän määrittelyn mukaisesti: Big Data on pragmaattinen datan analysoinnin muutos, jossa käytetään hyvin tunnettuja kaavoja, joiden avulla tietoon piilotetut relaatiot ja yhteydet saadaan esille. Näitä löydettyjä yhteyksiä voidaan hyödyntää korkeamman tiedon johtamisessa ja johdetun arvon käyttämisessä osana päätöksentekoprosessia. Tämä voidaan nähdä radikaalisena muutoksena yrityksien toimintatavoissa, joissa yritys pyrkii tuottamaan lisäarvoa itse keräämänsä ja tuottamansa tiedon pohjalta. Yritysten suurin ongelma ei ole enää päättää pitäisikö sen esimerkiksi julkaista uusi tuote markkinoille, vaan hyödyntää jo olemassa olevaa tietoa ja löytää ratkaisut sen pohjalta.

2.2 Big Datan kuudes V – Visualisointi

Yhä kasvavien datamäärien edessä, tiedon kuvaaminen ja esittäminen muodostuu yhä suuremmaksi ongelmaksi. Miten esittää tulokset ja päätökset selkeämmin? Useat mediat ja alan kolumneja kirjoittavat henkilöt ovatkin esittäneet datan visualisoinnin yhtenä Big Dataa kuvaavista V-määreistä ja pitävät sitä avaintekijänä Big Datan -aikakaudella [McNulty 2014, Khan & Khan 2011]. Visualisoinnin ongelmaa lähestyttäessä, suurimmiksi rajoittaviksi tekijöiksi nousevat ihmisen rajoittuneet kognitiiviset kyvyt käsitellä laajaa visualisointia sekä näyttöjen rajoitetut koot. Monien mielestä visualisoinnin tehokas käyttö onkin ainoa tapa siihen miten Big Datasta saadaan sen tuoma arvo esille ja tuotua se kaikkien käyttöön. [Wang et. al, 2015, NGrain 2013]. Myös Intelin [2013] mukaan jokaisen yrityksen, joka haluaa saada lisäarvoa Big Datasta, tulisi kääntää katseensa ensisijaisesti datan visualisointia edistäviin työkaluihin. Aikaisemmassa luvussa esiteltyjen kolmen V:n sekä viiden V:n mallien pohjalta, Intel esittääkin näitä tekijöitä yhdistävää mallia (Kuva 2), jossa tiedon visualisoinnilla on merkittävä rooli osana arvonluontiprosessia. Visualisoimalla tieto ymmärrettävään muotoon on arvoa luovien päätösten ja huomioiden tekeminen datan pohjalta mahdollista kaikkien työntekijöiden osalta. Big Datan onnistuneen visualisoinnin merkitys ja visualisoinnin mahdollistaminen toimivatkin päätekijöinä tässä tutkimuksessa.



*Kuva 2. Intelin esittämä neljän V:n malli korostaen datan visualisointia [Intel, 2013].*

Rijmenamin [2016] artikkelin mukaan ensimmäinen Big Dataa vaikuttavasti visualisoiva projekti oli Harrisonin ja Römhildin vuonna 2007 luoma visuaalinen kuvaus (Kuva 3) raamatun 63 779 sisäisestä ristiviittauksesta, jaettuna X-akselilla raamatun eri lukujen mukaisesti. Tässäkin tapauksessa sovellettua datajoukkoa voidaan kooltaan vielä pitää hyvinkin maltillisena Big Datan maailmassa, sisällön rajoittuessa hyvinkin alle sataan tuhanteen tietoalkioon. Harrisonin ja Römhildin esimerkistä voidaan myös todeta, että tässä käytetyn datajoukon sisältö oli hyvin tunnettu, jolloin joukon sisäiset tuntemattomat tekijät eivät pääse häiritsemään luotettavan visualisoinnin tekemistä. He antoivat kuitenkin työllään suuntaa sille, miten vaikuttavia visualisointeja voidaan luoda normaalia suurempien tietojoukkojen pohjalta.



*Kuva 3. Harrisonin ja Römhildin [2007] tekemä visualisointi raamatun tekstien sisäisistä viittauksista.*

Rijmenami [2016] myös toteaa artikkelissaan, että visualisointi ei teknologisesti olisi vaikein prosessillinen osuus Big Datan hyödyntämisessä, mutta se on haasteellisin osuus kokonaisuuden onnistumisen kannalta. Kompleksisen ja kvantitatiivista sekä kvalitatiivista sisältöä hyödyntävän tarinan esittäminen pelkästään graafien avulla vaatii uudenlaista ja kompleksisempaa näkökulmaa asiaan.

Big Datan lähteiden suoraviivainen visualisoiminen ei useinkaan ole mahdollista tai tehokasta ilman ennakkoanalyysia [Wang et. al*,* 2015]. IBM:n mukaan parhain lopputulos saavutetaan, kun liiketoiminnan analytiikan ratkaisut yhdistetään osaksi visualisointia [Keahey, 2013]. Käytännössä ja yksinkertaisuudessaan tämä IBM:n kohdalla tarkoittaa *Rapidly Adaptive Visualization Enginen* (RAVE) hyödyntämistä valmiin data joukon visualisoimisessa, johon käyttäjä on jo valmiiksi tehnyt datan rakenteen määrittelyn. Tällöin ei voida sanoa, että nämä ratkaisut edustaisivat visualisoinnissa aikaisemmasta eroavaa suuntausta tiedon visualisoinnin osalta, muuta kuin siltä osin, että järjestelmä pystyy käsittelemään isompia datamääriä. Käytännössä siis järjestelmälle tulee yhä määrittää spesifisesti sille annetun tiedon sisältö, jolloin järjestelmä ei pysty itsenäisesti mukauttamaan visualisointia tiedon sisällön muuttuessa. Tiedon ja Big Datan visualisointiin liittyviä tekijöitä tullaan käsittelemään vielä tarkemmin tutkimuksen Kappaleessa 3.

2.3 Big datan hyödyt ja ongelmat

Yritykset ja valtioiden toimijat näkevät Big Datassa suuren mahdollisuuden. Yhdysvaltain Valkoinen talo esitti jo vuonna 2012 julkaisemassaan tukirahoituskampanjassaan digitaalisen tiedon hyödyntämisen mahdollisuutena ratkaista valtion isoimmat ongelmat energiateollisuuteen, terveydenhuoltoon ja maan puolustukseen liittyen [Office of Science and Technology Policy, 2012]. Valkoisen talon tukirahoituskampanjassa tuodaan esille myös rahalliset avustukset visuaalisten ja graafisten tekniikoiden toteuttamiseen ja kehittämiseen liittyen kompleksista dataa varten. Erityisesti tarve nopeasti reagoivaan, visuaaliseen ja päätöksentekoa helpottavaan työkaluun tuodaan esille erilaisissa maan puolustukseen liittyvissä skenaarioissa ja järjestelmätarpeissa.

Big Datan hyödyntämisen on todettu tarjoavan paremman yleiskuvan yrityksen tuotantoprosessista (*supply chain*) ja johtavan asiakkaiden parempaan ymmärtämiseen, uusiin innovaatioihin [Monino, 2016] ja tuotantoprosessin optimoimiseen [Klein & Gorton, 2015]. Yritykset, jotka hyödyntävät Big Dataa, on arvioitu saavan merkittävää etumatkaa kilpailijoihin nähden heidän pystyessä ennakoimaan markkinan toimintaa tehokkaammin. Rajaraman [2016] antaa esimerkin laajasta kahvilaketjusta, jolla on useita liikkeitä ympäri Yhdysvaltoja. Kahvila lanseerasi uuden kahvisekoituksen ja turvautui sosiaalisen median kautta kerättyyn Big Dataan heidän analysoidessaan kuluttajakertomuksia uuden kahvilaadun suhteen. Julkistuspäivänä iltapäivään mennessä, yritys oli pystynyt seulomaan suurimmaksi ongelmaksi korkean hinnan ja muuttaneet tätä dynaamisesti ensimmäisen myyntipäivän aikana. Reaaliaikaisessa seurannassa negatiiviset palautteet olivat hinnan suhteen nähty loppuneen. Voidaankin todeta, että kun yritys saa prosessinsa sekä järjestelmänsä reaaliaikaista (*Velocity*) Big Dataa tukevaan tilaan, ja yritys pystyy tuottamaan kerätystä tiedosta merkityksellistä arvoa, yrityksen tehokkuus siirtyy uudelle tasolle.

Big Datan käyttöön liittyvät ongelmakohdat voidaan Akerkarin [2013] mukaan jakaa kolmeen ryhmään: Datan kompleksisuus (määrä, laajuus, eroavaisuus), Datan prosessointi (yhtäläisyyksien löytäminen, datan muokkaaminen ja analysoiminen) sekä Datan hallinta (yksityisyys, turvallisuus, eettisyys). Rajamankin [2016] nostaa tutkimuksessaan erityisesti ylös tilanteen datan hallinnasta: ”*Big Data on kuin kaksiteräinen miekka. Samalla kun se tarjoaa paljon hyödyllistä informaatiota kansalaisille, se johtaa myös yksityisyyden häviämiseen”.* Käyttäjät ovat harvoin tietoisia syöttäessään tietojaan eri järjestelmiin, että kuinka pitkälle tätä annettua tietoa voidaan yhä jalostaa, ja mitä kaikkea järjestelmät seuraavat käyttäjien toiminnassa. Käyttäjien seuraaminen verkkopalveluissa on sosiaalisen median palveluiden kautta luotu yhä helpommaksi. IP-osoitteiden, sosiaalisen median käyttäjätunnusten (joita yhä useammin käytetään palveluissa vaihtoehtoisena kirjautumistapana) ja laitteiden tunnistetietojen yhdistelyiden avulla, eri käyttäjien eri palveluiden tietoja voidaan yhdistellä suuremmaksi kokonaisuudeksi ja käyttäjän toiminnalliseksi kartastoksi. Erityisen huolestuttavan tilanteesta tekee se, että tietomurtojen kautta ihmisistä vuotaa yhä kasvavassa määrin erilaista dataa vääriin käsiin. Esimerkkinä Yahoo-palvelun murtaminen [Yahoo, 2016], mitä on tähän mennessä pidetty yhtenä isoimpana tietoon tulleista tietomurroista.

Rajaraman [2016] myös lisää, että datan sisältöön ei tule sokeasti luottaa vielä datan analysoinnin jälkeenkään. Big Dataa käsiteltäessä ei voida turvautua siihen, että kun tietoa on kerätty massiivisesti, data ikään kuin automaattisesti paljastaisi sisältönsä ja sisäiset suhteensa. Tämä on Rajaramanin mukaan kaukana todellisuudesta, sillä harvoin on mahdollista saada käyttöönsä oikeasti kaikkea dataa. Rajaraman mainitsee esimerkkinä vaalien ennakkotulosten arvioinnin, sillä näissä otanta on pieni, joten Big Dataan pohjautuvaa ennakoivaa (*predictive*) analyysia ei voi luotettavasti tehdä. Toisena ongelmana nähdään dataan ja tuloksiin luottaminen sokeasti data analyysin jälkeen. Mahdollisuutena on, että analysointivaiheessa datan sisäiset korrelaatiot ja suhteet on muodostettu tai ymmärretty väärin, jolloin analyysiprosessin tulos on virheellinen.

Big Datan tietokantahallintajärjestelmän (Big Data Management System, BDMS) voidaan pitää yleisnimityksenä Big Datan hyödyntämisessä käytetylle järjestelmäratkaisulle. Järjestelmän tulee kyetä säilömään ja prosessoimaan petatavun kokoisia datajoukkoja ja niiden tulee olla arkkitehtuuriltaan skaalautuvia, jaoteltuja, tehokkaita sekä sietää virheitä. [Fernandez et al. 2014, Marcos et al. 2013]. Jatkuvan datamäärän kasvun vuoksi Big Data -pohjaisten järjestelmien haasteena on erityisesti skaalautuvuus. Järjestelmän tulee pystyä toimimaan myös pitkälle tulevaisuudessakin, vaikka tietosisältö olisi laajentunut moninkertaiseksi alkuperäisestä. Tämän lisäksi reaaliajassa toimivien järjestelmien tulee mukautua virheisiin, jossa esimerkiksi hajautetussa laskennassa käytetyistä koneista osa lopettaa toimintansa tai tietovarastoina käytetyt moduulit täyttyvät [Neves & Bernardino 2015]. Myös Klein ja Gorton [2015] alleviivaavat skaalautuvuuden ongelmaa ja toteavat, että Big Data järjestelmän käyttöönoton jälkeen, järjestelmään tulevien syötelähteiden määrän ja kasvun kontrolli häviää tyystin. Täten järjestelmälle ei voi etukäteen asettaa minkäänlaisia määrityksiä sen laajuudesta, sillä nämä voivat tulla hyvinkin nopeasti vastaan. Siksi järjestelmän tilaa tulisi jatkuvasti monitoroida, ja seurata järjestelmän sisäisiä trendejä. Tällöin kattavan seurannan kautta pystytään ennakkoon vaikuttamaan järjestelmässä tapahtuviin muutoksiin tekemällä muutoksia hajautetussa järjestelmä- ja palvelinympäristöissä. [Klein & Gorton 2015]. Tarkemmin Big Data järjestelmien ongelmiin palataan vielä Kappaleessa 3, jossa käydään läpi erityisesti virtuaalitodellisuutta hyödyntävien järjestelmien luomat haasteet.

On arvioitu, että vuonna 2020 esineiden internet (*Internet of Things, IoT*) tulee yksistään tuottamaan 4 zettatavun verran dataa yhden vuoden aikana. [Turner et al. 2014]. Tällöin syötelaitteiden ja niiden luoma tietomäärä asettaa yhä kasvavan ongelman Big Data järjestelmien skaalautuvuudelle. Hajautettuun verkkoon tulee lisätä dynaamisesti uusia prosessoivia laitteita ja datan käyttö tulee jakaa näille tasaisesti. Watsonin [2014] mukaan on vaikea luoda yhtä geneeristä alustaa Big Dataa varten, jonka organisaatiot voisivat helposti ja nopeasti ottaa käyttöönsä. Usein tämä vaatiikin laajempaa järjestelmien kustomointia riippuen järjestelmän ympäristöstä ja järjestelmään ajetun tiedon ominaisuuksista. Big Datan trendiin ja käsitteeseen kuuluu vahvasti myös pilvipalvelut, jotka ovat alkaneet voimakkaasti yleistyä, mikä avaa tehokkaampia mahdollisuuksia isojen data määrien hallitsemiseen [Teräs & Raghunathan 2015]. Pilvipalveluiden myötä tiedot siirtyvät käyttäjien lokaaleilta kovalevyiltä pilvipalveluiden servereille, jolloin saatavilla olevan tiedon määrä kasvaa ja tiedon käyttö helpottuu.

3. TIEDON VISUALISOIMINEN

Tiedon visualisoinnin kappaleessa annetaan aluksi yleiskuvaus tieteellisen visualisoinnin alasta. Tieteellisen visualisoinnin perusta ja määritelmä käydään lävitse. Tämän jälkeen kuvataan visualisoinnin käyttötapauksia ja mahdollisuuksia esimerkkien avulla. Määrittelyiden jälkeen esitellään visualisoinnin tekniikoita ja menetelmiä. Näitä menetelmiä arvioidaan erityisesti virtuaalitodellisuuden ja Big Datan käyttöön soveltuvuuden näkökulmista. Lopuksi tehdään yhteenveto parhaimmin virtuaalitodellisuuteen ja Big Datan käyttöön soveltuvimmista visualisointitekniikoista.

3.1 Visualisointi yleisesti

Visualisoinnilla tarkoitetaan jonkin asian tekemistä havainnoitavaksi näköaistin avulla.Erityisen tärkeää visualisoinnissa on muistaa ihmisten sisäisten mallien muodostumisen tukeminen. Donalekin ja kumppaneidenkin [2014] mukaan visualisointi on pääväylä kvantitatiivisen tietosisällön ja ihmisen tietoisuuden välillä. Ajatuksen pohjana on, että ihminen ei kykene täysin ymmärtämään asioita, joita ihminen ei pysty jollakin tavalla visualisoimaa. Ihmisille onkin kehittynyt merkittävä taito kaavojen ja yhteyksien tunnistamiseen. Tämän kautta taito löytää tietoa datan ohjaamassa (*data-driven*) tieteessä pohjautuu merkittävästi onnistuneeseen datan visualisointiin, jossa datan tutkiminen olisi tehokasta ja joustavaa.

Visualisointi nähdään monialaisena käsitteenä sisältäen tekniikoita tietokonegrafiikasta, kuvankäsittelystä, konenäöstä, tietokoneavusteisesta opetuksesta, geometrisesta mallinnuksesta, psykologiasta ja käyttöliittymäsuunnittelusta [Haber & McNabb, 1990]. Monialaisuuden takia visualisoinnin on nähty edistäneen kehitystä monilla eri tutkimusaloilla [Johnson et al., 2005]. Visualisoinnin ala jaetaan perinteisesti kuitenkin kahteen eri osioon: Tieteelliseen visualisointiin ja Informaation visualisointiin.

Tieteellisessä visualisoinnissa pyritään yhdistämään ihmisen kognitio osaksi tietokoneita ja niiden tuottamaa grafiikkaa. Käytännössä tieteellisen visualisoinnin ala keskittyy tieteellisten tutkimustulosten liittämiseen osaksi reaalimaailman prosesseja. Tutkimusalana visualisoinnissa pyritään helpottamaan tiedon ymmärtämisen prosessia ohjelmistotyökalujen avulla, jotka tarjoavat staattisia tai interaktiivisia visualisoinnin esityksiä. [Johnson et al., 2005].

Informaation visualisointi syntyi omana haaranaan ihmisen ja tietokoneen välisen vuorovaikutuksen tutkimusalan *(Human Computer Interaction, HCI*) syntyessä 1980 -luvun lopulla. Informaation visualisoinnin tutkimusalan tarkoituksena on myös tiedon ymmärtämisen ja käsittelyn helpottaminen, mutta pääpaino tutkimuksessa on erityisesti mentaalimallien ja tietokonegrafiikan hyödyntämisessä.

Brown ja kumppanit [1996] näkevät, että visualisoinnilla on kolme keskeistä päämäärää: tutkiminen, analysointi ja esittäminen. Visualisointia siis käytetään olemassa olevan data joukon tarkasteluun ja mielenkiintoisten rakenteiden hahmottamiseen jatkoanalysointia varten. Visualisointia voidaan yhä käyttää hypoteesien ja saatujen tulosten varmentamiseen sekä esittämiseen esimerkiksi muille kollegoille tai julkiselle yleisölle. Visualisointi on kuitenkin harvoin eriytetty, itsenäinen prosessinsa: se on usein tarpeellinen, mutta ei yksistään riittävä tapa ongelmien ratkaisemiseen. Tiedon visualisoiminen vaatiikin rinnalleen yhä vahvemmin analyyttisia välineitä ja tekniikoita kuten tilastotiedettä, tiedonlouhintaa ja kuvan prosessoimista. Johnson ja kumppanit [2005] esittävätkin raportissaan visualisointitutkimuksen yhtenä suurimmista ongelmista alan mukautuvuuden kaikkiin muihin aloihin, joiden ongelmia pyritään visualisoinnin avulla selvittämään käyttäen yhä laajempaa määrää tietoa. Tämän haasteen pohjalta onkin esitelty uusi poikkitieteellinen käsite Tieteellisen ja Informaation visualisoinnin rinnalle: Visuaalinen analytiikka (*Visual Analytics*). Visuaalinen analytiikka koostaa aikaisempia lähestymistapoja yhteen, mutta painottaa erityisesti visuaalisten teknologioiden sekä työkalujen kehittämistä ja hyödyntämistä.

Tässä tutkielmassa visuaalisuutta käydään läpi erityisesti analyyttisen puolen osalta, sillä työssä pyritään kartoittamaan ratkaisuja datan helpompaa analysoimista varten hyödyntämällä uusia virtuaalitodellisuuden työkaluja. Informaation visualisointiin pohjautuvasti, työssä pyritään löytämään mahdollisimman hyvin käyttäjien mentaalimalleja tukeva ja virtuaalitodellisuuden mahdollisuuksia hyödyntävä visualisointitapa Big Dataksi luokiteltavan tiedon esittämistä varten. Tutkielmassa virtuaalitodellisuuden teknologia nähdään uutena työkaluna ja sitä hyödyntämällä luotuja visualisointeja pyritään vertaamaan visualisointeihin normaalissa työasemaympäristössä. Ensisijaisena päämääränä tutkimuksessa on selvittää, lisääkö virtuaalitodellisuus informaation visualisoinnin tehokkuutta ja ymmärrystä verrattuna perinteiseen työasemaympäristöön ja 2D-visualisointeihin.

3.2 Visualisoinnin tyypit

Tiedon visualisoinnin menetelmät ovat menneet digitalisaation myötä vauhdilla eteenpäin viime vuosikymmenien aikana. Alkuperäisesti tiedon visualisointia hyödynnettiin karttatieteissä ja liiketoimintaan pohjautuvan tilastollisen tiedon esittämisessä [Few & Edge, 2007]. Nykykäsitteen mukaisen tiedon visualisoinnin katsotaan toteutuneen ensimmäisen kerran William Playfairin toimesta, kun hän esitteli teoksessaan *Commercial and Political Atlas* [1786] Englannin kaupankäyntiin pohjautuvaa tilastoa graafien avulla. 2000–luvun aikana tiedon visualisoinnissa on keskitytty erityisesti edistämään tiedon sisällön interaktiivisuutta ja erilaisia vuorovaikutustapoja visualisoidun sisällön kanssa [Kehrer et al, 2012]. Interaktiivisuuden merkitys korostuu, kun visualisointi toteutetaan 3D-mallinnettuna, jolloin vuorovaikutuksessa hyödynnettävät mahdollisuudet kasvavat dimensioiden määrän kasvaessa. Nykyään tiedon visualisoinnissa ongelmia tuottaa datan suuri määrä ja visualisoinnin skaalautuvuus datan määrän mukaisesti. Skaalautuvuuden ongelma määriteltiin Kappaleessa 2.1. Tämän mukaisesti skaalautuminen on määritelty yhdeksi isoimmista ongelmista Big Datan hyödyntämisessä. Lisäksi Chen & Zhang [2014] toteavat, että oikean visualisointitavan valinta on merkittävin tekijä Big Datan visualisointiprosessissa.

Perinteisimmiksi tiedon visualisointitavoiksi voidaan luokitella: taulukot, piirakkakaaviot, pylväsdiagrammit, viivakaaviot, pisteparvet, vuokaaviot ja esimerkiksi aikajanat. Perinteiset visualisointimenetelmät muuttuvat kuitenkin tehottomiksi ja epäselviksi datan määrän kasvaessa. Esimerkiksi piirakkakaaviossa hyödynnetään värikoodausta eri osien erottamiseen toisistaan. Big Datan kohdalla tietueiden määrä on kuitenkin niin laaja, että toisistaan erottuvat värikoodaukset loppuvat kesken. Myös Ware [2004] toteaa kirjassaan, että ihmisen visuaalinen työmuisti on rajoittunut kolmesta viiteen kohteeseen kerralla, jolloin värikoodausta tulisi hyödyntää vain pienissä visualisoinneissa. Yksi suosituimmista monimuuttuja-analyysissä käytetyistä visualisointitekniikoista on pisteparvitaulukko *(Scatter Plot Matrix, SPLOM*) [Card et al. 1990]. Pisteparvet auttavat segmenttien, raja-arvojen, trendien ja korrelaatioiden löytämisessä, mutta laajaa moniulotteista dataa kuvattaessa joudutaan hyödyntämään useita eri arvoja kuvaavia pisteparvitaulukoita. Toinen ongelma pisteparven kohdalla esiintyy, kun datapisteitä on niin paljon, että pisteet alkavat piirtymään toistensa päälle (*overplotting*). Visualisoinnin sotkeutuminen (*clutter, visual clotting*) on päälle piirtymisestä juontuva ongelma, jolloin ei voida enää hahmottaa datapisteiden sijaintia ja raja-alueita visualisoinnin sisällä [Du et al. 2016, Fisher 2016]. Visuaalisuuden heikkenemisen ongelma on esitettynä Kuvassa 3*.*



*Kuva 3. Visuaalisuuden heikkenemisen ongelma (visual clotting) kuvattuna 3D-pisteparvessa. [Du et al. 2016].*

Big Dataa kuvattaessa tarpeet visualisoinnille säilyvät hyvin samanlaisina kuin normaalissakin tilanteissa, mutta näiden tarpeiden saavuttaminen vaatii normaalia enemmän työtä. Olshannikova ja kumppanit [2015] määrittävät tutkimuksessaan big datan visualisoinnin merkittävimmiksi tekijöiksi seuraavat: piilotettujen yhteyksien tai anomalioiden identifiointi, spesifisten arvojen etsimisen joustavuus, eri arvojen kvantitatiivinen vertailu ja käyttäjän reaaliaikainen vuorovaikutus visualisoinnin kanssa. Näiden tavoitteiden saavuttaminen kuitenkin vaikeutuu käsiteltäessä huomattavasti isompia datajoukkoja, joihin aikaisemmin esitetyt, perinteiset visualisointitekniikat, taipuvat huonosti. Perinteisempien visualisointitekniikoiden heikon Big Data soveltuvuuden takia, seuraavaksi käydään lävitse normaalista poikkeavampia visualisointitapoja, jotka tukevat erityisesti kompleksisen, monia ulottuvuuksia sisältävän datan käyttöä. Visualisointitekniikoista arvioidaan niiden soveltuvuutta isojen tietomäärien visualisoimiseksi ja esitellään tekniikoiden vahvuudet ja heikkoudet.

3.2.1 Parallel coordinates

*Rinnakkaisten koordinaattien* tekniikkaa käytetään yksittäisen dataelementin piirtämiseen useiden dimensioiden välillä. Täten tekniikka soveltuu erityisen hyvin moniulotteisen datan esittämiseen ja on erittäin laajasti käytetty. [Wang *et al.* 2015]. Pääperiaatteena on esittää jokaisen tietueen arvot sarjana vierekkäisiä akseleita ja jokainen arvo on linkitetty osaksi omaa akseliaan. Visualisointityyli voidaan nähdä yksinkertaisemmillaan kuin taulukkona, jonka rivien sarakkeiden välille on vedetty yhdistävät viivat. Tämän takia rinnakkaisten koordinaattien avulla voidaan visualisoida useita arvoja sisältävät tietueet yhdessä kuvaajassa. Myös rinnakkaisten koordinaattien visualisointi sisältää saman ongelman kuin pisteparvi: Tietomäärän kasvaessa viivat alkavat vahvasti piirtymään toistensa päälle, jolloin visualisoinnista on enää vaikea erottaa mitään. Täten visualisointitekniikkaa on vaikea hyödyntää isoille data joukolle, jonka tietueiden määrä nousee yli kahden tuhannen.



*Kuva 4. D3-kirjaston luoma esimerkki rinnakkaisten koordinaattien visualisoinnista, jossa on kuvattuna automallien tietoja.*

3.2.2 Star coordinates

Tähtikoordinaatit ovat yksinkertainen, tehokas ja hyvin tunnettu tiedon interaktiotapa moniulotteisen datan visualisoimiseen. Yleisimmin tekniikkaa käytetään tutkimuksellisiin tarkoituksiin kuten klusterien analysoimiseen, poikkeavien havaintojen suodattamiseen ja trendien havaitsemiseen [Rubio-Sanchez & Sanchez, 2014]. Tähtikoordinaattien käytön tarkoituksena onkin luoda helposti ymmärrettäviä moniulotteisia visualisointeja, jotka tukevat data joukon sisällön hahmottamisprosessia. Täten päämäärä ei ole numeerinen analyysi vaan yleiskäsityksen luominen tiedon sisällöstä [Kandogan 2000, 2001].

Tähtikoordinaattitekniikan ideana on järjestää koordinaattiakselit ympyrän sisälle, jossa jokainen samanpituinen akseli omaa lähtökohtaisesti yhtä suren kulman ympyrän keskustan ja akselin välillä. Chenin [2014] mukaan tähtikoordinaattitekniikkaa voidaan hyödyntää jopa miljardien eri tietueiden yhtä aikaiseen kuvaamiseen, sillä visualisointi tukee päällekkäin menevien tietojen aggregointia syvyystiedon muodossa toisin kuin esimerkiksi rinnakkaiset koordinaatit. Lisäksi Chen toteaa tutkimuksessaan, että tähtikoordinaatit soveltuvat erityisen hyvin Big Datan visualisointiin, sillä visualisoinnissa ei tarvitse laskea tietueparien (*pairwise*) välisiä etäisyyksiä, vaan tieto etäisyyksistä säilötään visualisoinnin pohjalla toimivaan malliin. Tämä erottelu mahdollistaa esimerkiksi aikaisemmin mainitun syvyystietojen hyödyntämisen visualisoinnissa.

Tähtikoordinaatteihin pohjautuvaa klusterianalyysia ja validointia on hyödynnetty esimerkiksi Kandoganin tutkimuksessa [2001], VISTA -järjestelmässä [Chen & Liu, 2004] sekä Longin & Linsenin [2011] moniulotteista dataa käsittelevässä tutkimuksessa. Tähtikoordinaattien hyödyntäminen vaatii myös käyttäjän interaktiota ja harvoin visualisointi on heti alussa käyttäjälle arvoa tuottava ilman käyttäjän tekemiä päätöksiä ja analyysiä. Tärkeimmäksi näistä voidaan lukea alpha arvon mukauttaminen. Arvon muuttaminen tapahtuu skaalaamalla akselien pituutta, jonka avulla pystytään lisäämään tai vähentämään eri ulottuvuuksien painoarvoa visualisoinnissa. Muihin yleisimpiin vuorovaikutustapoihin lukeutuvat esimerkiksi arvoalueiden antaminen, tietueiden valitseminen ja arvojen välisten korrelaatiopainotusten muuttaminen [Kandogan 2000]. Muokkauksen jälkeen visualisointi joudutaan piirtämään kokonaan uudelleen. Täten käyttäjän tekemien, jatkuvien uudelleenmääritysten takia, vuorovaikutusprosesseihin tulee kiinnittää erityistä huomiota Big Dataa hyödyntävissä järjestelmissä, jotta muutokset pystytään prosessoimaan nopeasti heikentämättä käyttökokemusta.

Tähtikoordinaattien tukiessa moniulotteista dataa ja soveltuessa myös massiivisten tietojoukkojen esittämiseen, tähtikoordinaatit voidaan nähdä kattavimpana visualisointivaihtoehtona Big Dataa varten, sillä teoriassa data joukon suuruudella ei ole vaikutusta visualisoinnin onnistumiseen. Lisäksi Kandoganin [2000, 2001] tutkimusten mukaan, tähtikoordinaattien tekniikka soveltuu erityisesti data-analyysin ensimmäisiin vaiheisiin, joissa pyritään hahmottamaan tietojoukon yleisiä ominaisuuksia. Tämä prosessinvaihe tulee olemaan painotettuna myös tämän tutkimuksen virtuaalitodellisuus-sovellutuksessa.

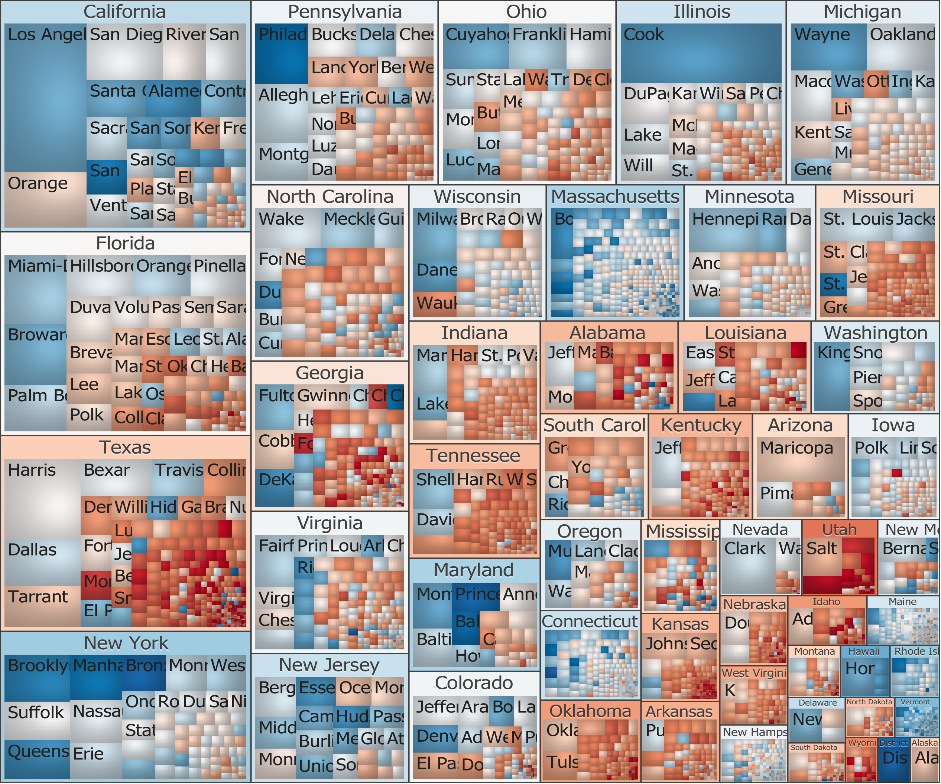


*Kuva 5. Long & Linsen [2011] kuvasivat 10-ulotteista (vas.) ja 20-ulotteista dataa 3D-muodossa tähtikoordinaattien avulla.*

3.2.3 Tree map

Klassinen puukartta tarjoaa mahdollisuuden aggregoida tietoa datajoukon hierarkian ylemmillä tasoilla ja tarjoaa hierarkian avulla yhä tarkempia näkymiä tiedon sisällöstä [Shneiderman 1992]. Puukartoilla voidaan esittää hierarkista dataa, jaoteltuna sisäkkäisiin suorakulmioihin. Jokainen tietoelementti toimii puun yhtenä haarana, jolle annetaan suorakulmio, mikä sisältää elementin sisältämät arvot jaettuna uusiin, pienempiin suorakulmioihin. Puukartoissa usein hyödynnetään tiilien kokojen ja värien korrelaatiota, jolloin tiedon sisältö ja merkittävyys on käyttäjälle helpommin hahmotettavissa. Puukarttojen hyvänä puolena on lisäksi se, että visualisointi käyttää tehokkaasti tilaa hyödykseen ja mukautuu pienempäänkin tilaan.

Puukarttojen sisältäessä hyviä ominaisuuksia, se voitaisiin sovellettuna versiona nähdä hyvänä vaihtoehtona tiedon visualisoinnille virtuaalitodellisuuden ympäristössä. Kuitenkin erityisesti Puukarttojen kohdalla, visualisointitekniikan hyödyntäminen riippuu yhä vahvemmin datan sisällöstä, sillä kuten mainittua, puukarttojen visualisointi nojaa vahvasti tiedon hierarkkisuuteen. Tietojoukon ollessa sisällöltään ja ulottuvuuksiensa osalta vahvasti heterogeenistä, tietojoukon sisäisistä korrelaatioista tulee epäluotettavia ja puukartaston luomasta visualisoinnista tulee hankalasti tulkittava.



*Kuva 6. Puukartta luotuna USA:n vuoden 2012 presidentinvaalien äänestystuloksen jakautumisesta.*

3.3 Visualisoinnin työkalut

Datan visualisointi tarkoittaa tiedon esittämistä järjestelmällisessä muodossa, sisältäen muuttujien ja yksikköjen tiedot [Khan & Khan 2011].

|  |  |
| --- | --- |
| **Edut** | **Prosentit (%)** |
| Parantunut päätöksenteko | 77 |
| Parempi ad-hoc data-analyysi | 43 |
| Parantunut yhteistyö ja tiedon jakaminen | 41 |
| Itsepalvelumahdollisuuksien tarjoaminen loppukäyttäjille | 36 |
| Kasvanut ROI (Return on investment) | 34 |
| Aikasäästöt | 20 |
| Vähentynyt IT-kuorma | 15 |

*Taulukko 1. Datan visualisointityökalujen hyödyt [Sucharitha et al, 2014].*

Sucharitha ja kumppanit [2014] esittivät Big Dataa käsittelevässä kyselytutkimuksessaan visualisointityökalujen suurimmat hyödyt (Taulukko 1*).* Tutkimuksen tulosten pohjalta suurin osa vastaajista oli sitä mieltä, että tiedon visualisointia voidaan erityisesti käyttää osana tehokkaampaa päätöksentekoprosessia. Visualisointityökalujen käyttö tuo myös tiedon sisällön yhä laajemman segmentin käyttöön, kun tieto on muunnettu ymmärrettävään ja helpommin lähestyttävään muotoon. Tällöin visualisoinnit tarjoavat yrityksille mahdollisuuden havaita tuotteen, myyntien ja asiakkaiden välisiä korrelaatiosuhteita, jolloin kohdemarkkinointi tehostuu. Visualisointi tarjoaa myös työvälineen yrityksen toiminnan seurantaan, jota voidaan hyödyntää riskianalyysien tekemisessä.

Data määrien ja datan kompleksisuuden kasvaessa, yhä pidemmälle menevien tavoitteiden saavuttaminen työkalujen avulla kuitenkin vaikeutuu. Isoja tietojoukkoja käsiteltäessä on yleensä tukeuduttu tiedon tiivistämiseen käsittelyn helpottamiseksi, mutta Big Dataa tiivistettäessä ei voida suoraan määrittää, mikä osa tiedosta voidaan jättää pois. [Olshannikova et al. 2015]. Visualisointityökalujen haasteena onkin yhdistää abstrakti tieto osaksi reaalimaailmaa visuaalisen esityksen kautta. Näistä tekijöistä johtuen, visualisointityökalujen tulisi jatkossa tarjota yhä joustavampia ratkaisuita data joukon määrittelyyn ja analyysiin. Tämän lisäksi visualisointityökalujen esityksen tulisi myös täyttää ainakin seuraavat kolme vaatimusta: Ilmaisevuus (esitä vain tieto, jota data sisältää), tehokkuus (pohjautuen ihmisen kognitiiviseen havainnointiin) ja soveltuvuus (visualisoinnin hyöty-kustannus suhde) [Miksch & Aigner, 2014 ja Muller & Schumann 2003].

Datajoukkojen alati laajentuessa ja muuttuessa yhä kompleksisemmiksi tutkijat ovatkin alkaneet visualisoinneissa painottaa vuorovaikutustyökalujen tärkeyttä. Tällöin päätös sisällön rajaamisesta, tulkinnasta ja joukon sisäisten yhteyksien etsinnästä annetaan yhä vahvemmin käyttäjälle. Päämääränä on tarjota lähtökohtaisesti mahdollisimman paljon dataa käyttäjän näkyville ja tarjota mahdollisimman tehokkaat työkalut datan tulkintaan. Tällöin visualisoinnin käytöstä tulee joustavampaa, minkä tulisi tehostaa analysoinnin tekemistä. Wang ja kumppanit [2015] toteavatkin tutkimuksessaan, että käyttäjän osallistaminen ja interaktiivisten työkalujen käyttö on hyvin tärkeää, sillä staattiset visualisoinnit eivät tee autuaaksi ja ovat huomattavasti tehottomampia datajoukkoa analysoitaessa. Visualisoinneissa tulisikin nojautua ihmisen kognitiiviseen kykyyn havaita visuaalisia malleja ja siirtää kriittinen ajattelu käyttäjän vastuulle. Khan & Khan [2011] nimeävät tutkimuksessaan vuorovaikutteisen visualisoinnin vaiheet:

1. *Valitseminen:* Käyttäjällä tulee ollamahdollisuus valita yksittäinen tietue, alijoukko tai koko datajoukko oman kiinnostuksen mukaisesti.
2. *Linkitysten tekeminen:* Mahdollisuus linkittää tietueita toisiinsa ja vertailla niiden sisältämiä arvoja eri näkymissä.
3. *Suodattaminen:* Auttaa käyttäjää muuttamaan esillä olevan tiedon määrää ja auttaa keskittämään fokuksen niihin elementteihin, joista ollaan kiinnostuneita.
4. *Uudelleen järjestäminen:* Spatiaalisen näkymän ollessa tärkein tapa visuaalisessa havainnoinnissa, tulee käyttäjällä olla mahdollisuus muuttaa tiedon asettelutapaa ja näkymää uusien näkökulmien saamiseksi.

3.4 Visualisoinnin prosessi

Visualisointityökalujen haasteiden ja vaatimusten esittelyn jälkeen käydään läpi yleisluontoinen prosessimalli visualisoinnin toteuttamiseksi. Tietotekninen prosessi raa'an datan johtamisesta visualisoinniksi (*visualization pipeline*) noudattaa vahvasti edellisessä kappaleessa Khan & Khanin [2011] käyttäjän näkökulmasta tekemää vuorovaikutteisen visualisoinnin prosessia. Järjestelmätason visualisoinnin prosessi on määritelty koostuvan seuraavista vaiheista: Tiedon analysoiminen, suodattaminen, kartoittaminen ja kuvantaminen [InfoVis].

1. *Tiedon analysoinnissa* data valmistellaan visualisoimista varten esimerkiksi poistamalla puuttuvat tai virheelliseksi määritetyt arvot tai suodattamalla osa ei-halutuista arvoista pois.
2. *Suodattamisessa* valitaan halutut osat tietojoukosta visualisointia varten usein käyttäjän toimesta.
3. *Kartoittamisessa* data, josta ollaan kiinnostuneita, liitetään osaksi geometrisiä primitiivejä, (esimerkiksi pisteet ja viivat) ja niiden ominaisuuksia (väri, sijainti, koko).
4. *Kuvantamisessa* aikaisempien vaiheiden pohjalta muodostettu geometrinen data muunnetaan kuvalliseen, visualisoituun muotoon.



*Kuva 7. Visualisoinnin prosessi kuvattuna..*

Perinteistä visualisoinnin prosessia hyödyntäviä järjestelmiä on kehitetty useita vuosien varrella. Nykyään visualisoitavan datan määrä asettaa näille kuitenkin haasteensa, sillä tiedon käsittely, suodattaminen ja yhteen liittäminen suoritetaan offline -tilassa paikallisesti, välittämättä kustannuksista. Datan määrän kasvaessa ongelmat kasvavat, kun yhä enemmän siirretään tietoa säilövän moduulin (tietokanta) ja kuvantamisen suorittavan moduulin (käyttöliittymä) välillä (*client – server malli*) [Vo et al*.,* 2011]. Myös Moreland [2013] toteaa tutkimuksessaan, että nykyiset, ahneet visualisoinnin algoritmit on suunnattu tarjoamaan lyhytaikaista laskentaa isolle datajoukolle. Toisena ongelmana Moreland näkee visualisointijärjestelmien huonon skaalautuvuuden jatkuvaan ja dynaamiseen datajoukkojen käsittelyyn. Aikaisempien visualisointityökalut ovat muuntautuneet huonosti tukemaan uusia ja kompleksisempia tietorakenteita, joten ratkaisuja on lähdetty hakemaan muualta. Suurten datajoukkojen käsittelyä ja visualisointia varten onkin kasvamassa määrin alettu käyttämään *MapReduce* – arkkitehtuuria. MapReduce -arkkitehtuuri onkin suunniteltu suurten data määrien nopeaa käsittelyä varten.

3.5 MapReduce

*MapReduce* on yleiskäyttöinen ja kevyt viitekehys, joka on kehitetty erityisesti tiedon rinnakkaiseen prosessointiin hajautetussa järjestelmäympäristössä. *MapReduce* on osa *Apache Hadoopin* avoimeen lähdekoodiin pohjautuvaa kirjastoa, jonka tarkoituksena on mahdollistaa isojen tietomassojen käsittely hajauttamalla tiedon prosessointi pilvipalveluympäristössä useiden laitteiden hoidettavaksi. *MapReduce* perustuu kahden operaation abstraktioon:

* *Map:* Prosessoi saamansa avain/arvo parit ja tuottaa näistä nolla tai enemmän avain/arvo pareja tuloksena.
* *Reduce:* Kutsutaan kerran jokaisen uniikin avaimen kohdalla. Funktio iteroi kaikkien arvojen lävitse, jotka jakavat saman avaimen ja tuottaa tuloksena nollan tai nollaa isomman arvon.



*Kuva 8. Simuloitu kuvaus MapReduce – funktion toiminnasta.*

*(*[*http://www.edureka.co/big-data-and-hadoop-course-curriculum*](http://www.edureka.co/big-data-and-hadoop-course-curriculum)*)*

*MapReducen* suoritus alkaa *Map* -vaiheella, jossa jokainen avain/arvo pari luodaan annetun syötteen pohjalta. Tämän jälkeen *Shuffle* -vaiheessa edellisen vaiheen avain/arvo parit ryhmitellään avaimen mukaisesti (saman avaimen omaavat samaan ryhmään). Lopuksi *Reduce* -operaation avulla avain/arvo pareista koostetaan lopullinen tulos yhdistämällä arvot yhtenevän avaimen alaisuuteen. Tarkemmin operaation suoritus on kuvattu *Kuvassa 4.* missä simuloidaan sanojen määrän simulointia saadusta syötteestä.

Hadoopin ja MapReducen käytön avulla vältytään myös hajautetun ohjelmasuorituksen, datan hajauttamisen ja jakamisen sekä virheistä palautumisen ongelmiin varautumiselta, mitkä tarjotaan suoraan Hadoop – kirjaston toimesta. Nämä yhdessä tekevät MapReducesta yksinkertaisen, mutta tehokkaan vaihtoehdon suurten datajoukkojen käsittelyyn, yhdistettynä vielä järjestelmän riippumattomuuden luettavaksi annetun datan määrästä [Vo et al., 2011]. Vo ja kumppanit [2011] myös huomioivat tutkimuksessaan, että yleisiä visualisointiin käytettyjä algoritmeja voidaan luonnollisesti kuvata *MapReducen* käyttämällä abstraktiolla ja luoda täten yksinkertaisia sekä erittäin hyvin skaalautuvia järjestelmiä.

4. VIRTUAALITODELLISUUDEN HYÖDYNTÄMINEN

4.1 Virtuaalitodellisuus aikaisemmin

Tieteellisessä tutkimuksessa virtuaalitodellisuus on jo pidempään nähty tehokkaana alustana ihmisen ja teknologian välisessä vuorovaikutuksessa [Burdea & Coiffet, 2003]. Kim [2005] määritteli aikoinaan tutkimuksessaan virtuaalitodellisuuden järjestelmäksi, joka takaa saumattoman käyttökokemuksen ja koostuu useiden näyttöjen kokoonpanosta. Näiden tehokkaiden ja innovatiivisten laitteiden onkin todettu tarjoavan työkalun moniulotteisen ja kollaboratiivisen datan visualisoimiseen. Wangin ja kumppaneiden [2015] mukaan virtuaalitodellisuudella pystytään erityisesti edistämään tiedon geometrian ymmärrystä (muotojen, kokojen merkitys) sekä hahmottamaan tiedon sisältö intuitiivisemmin tehokkaamman visualisoinnin kautta.

CAVE–ympäristöjä (*Cave Automatic Virtual Environment*) on jo pitkään hyödynnetty osana tieteellisiä tutkimuksia, joissa on haluttu hyödyntää virtuaalitodellisuuden elementtejä. CAVE pohjaiset järjestelmät ovat olleet hyvin kalliita ja niiden pystyttäminen on vaatinut paljon aikaa. Oculus Rift ja HTC Vive esimerkkeinä, tarjoavatkin nykyään halvemman ratkaisun hyödyntää virtuaalitodellisuutta eri toimialueilla ja sen mahdollisuudet on huomattu myös tutkimuspuolella. Ensimmäinen HMD-tyyppinen (Head Moundted Display) laite kehitettiin Sutherlandin [1968] toimesta ja Cox, Patterson sekä Thiebaux [Cox et al., 1997] lähettivät patentin liittyen äänen sekä eleiden hyödyntämiseen 3D-ympäristöissä. Cruz-Neira sekä kumppanit [Cruz-Neira et al., 1993] hyödynsivät jo aikaisessa vaiheessa CAVE-ympäristöä tutkimuksessaan ja Beck [2003] hyödynsi virtuaalitodellisuutta omassa kaupunkeja mallintavassa VRGIS-järjestelmässä, joka tunnetaan tutkimusalalla hyvin. Myöhemmin Foo ja kumppanit [2009] ottivat virtuaalitodellisuuden mukaan jo terveydenhuoltoonkin ja käyttivät virtuaalitodellisuutta osana endoskooppisten operaatioiden suunnittelua. Näiden lisäksi virtuaalitodellisuuden tutkimus on vahvasti pyrkinyt löytämään ratkaisuja vaikeista motorisista vammoista kärsivien ihmisten kuntoutukseen. Hyödynnetyn teknologian näkökulmasta katsottuna CAVE-tyyliset ratkaisut ovat tällä hetkellä tutkimuksissa korvautumassa HMD-pohjaisilla virtuaalitodellisuuden laitteilla, mikä on varsin ymmärrettävä trendi. Virtuaalitodellisuuden ala on kuitenkin hyvin uusi ja sen potentiaali on vielä nykyäänkin huonosti hyödynnetty. Tästä kertoo virtuaalitodellisuusalustoille tuotetun sisällön ja ratkaisujen niukkuus pelejä lukuun ottamatta. Tutkimuksellisessa mielessä tilanne on tietysti hyvä, koska se avaa mahdollisuuden pohtimaan, mitä kaikkea virtuaalitodellisuudella voitaisiin saada aikaan.

Virtuaalitodellisuuden esittelyn jälkeen, tässä työssä käydään yleisesti lävitse virtuaalitodellisuuden HMD-pohjaisen teknologian sisältö sekä teknologia ja keskitytään erityisesti HTC:n keittämiin *Vive*–virtuaalitodellisuuslaseihin, joita tullaan myös tässä tutkielmassa myöhemmin hyödyntämään.

Alan kehityksen myötä virtuaalitodellisuuslasit mahdollistavat jatkossa myös katseenseurannan, minkä hyödyntäminen on erityisen tärkeää. Tällä hetkellä virtuaalitodellisuuteen pohjautuvassa sisällöntuottamisessa on haasteena erityisesti käyttäjän vapaus katsoa mihin haluaa, jolloin kriittinen sisältö saattaa mennä ohitse. Tällöin katseenseurannan avulla voidaan tarjota käyttäjälle huomioita esitettävään sisältöön liittyen, jos käyttäjän huomio on kiinnittynyt toisaalle.

4.2 Big Datan visualisointijärjestelmät

Soveltuva tiedon visualisointi on Tengin ja kumppaneiden [2015] mukaan suurin helpottava tekijä Big Datan hyödyntämisessä ja analysoinnissa. Myös Plugfelder ja Helmut [2013] pitävät kattavaa tiedonvisualisointia vaatimuksena sille, että kerätty tieto saadaan valjastettua käyttöön ja saatettua myös vähemmän asiantuntevuutta omaavien henkilöiden käyttöön. Virtuaalitodellisuutta hyödyntäviä Big Datan visualisointijärjestelmiä ei ole kaupallisina versioina saatavilla, mutta aiheeseen liittyviä tutkimuksia ja tutkimuksellisia järjestelmiä on tehty muutamia. Seuraavaksi käydään läpi kolme, erityisesti Big Datan visualisointiin pohjautuvaa, virtuaalitodellisuuden järjestelmää ja arvioidaan niiden tutkimuksellista sisältöä tätä tutkimustyötä silmällä pitäen.

Donalekin ja kumppanien [2014] tekemää tutkimusta voidaan pitää ensimmäisenä työnä, jossa virtuaalitodellisuuden hyötyjä Big Datan visualisoinnissa on tutkimuksellisesti arvioitu. He kehittivät tutkimuksessaan iVIZ-visualisointijärjestelmän, joka on suunnattu käytettäväksi tiedon analysoimista varten virtuaalilasien tai suoraan selaimen kautta. Donalekin ja kumppaneiden [2014] tutkimuksen lähtökohta oli hyvin samanlainen kuin tässäkin tutkimuksessa: Hyödyntää virtuaalitodellisuutta yleisenä, abstrahoituna visualisointityökaluna, joka tarjoaisi mahdollisuuden silmäillä ja analysoida mitä tahansa tietoa. Donalekilla ja kumppaneilla [2014] oli selvä näkökulma tutkimuksessaan siitä, että algoritmit eivät vielä nykyään pysty löytämään piilotettuja kytköksiä tietojoukon sisältä vaan tulee hyödyntää primitiivisintä järjestelmää, eli käyttäjän havainnointikykyä. Tästä syystä iVIZ-järjestelmässä koko tietojoukko piirretään virtuaalimaailmaan ilman esisuodatusta. Järjestelmässä dataelementit kytketään osaksi XYZ-akselistoa, johon käyttäjä pystyy tekemään omia määrityksiään. Käyttäjälle annetaan mahdollisuus tutkia sisältöä virtuaalimaailmassa täysin vapaasti eri suodatusmekanismeja hyväksikäyttäen.



*Kuva 9. Donalekin ja kumppaneiden [2014] kehittämän iVIZ-järjestelmän perusnäkymä.*

Tutkimuksessa löydettiin viisi eri näkökulmaa miten tietoa voitaisiin esittää mahdollisimman kattavasti osana yhtä dataelementtiä:

* *XYZ –* Sijainti XYZ-koordinaatistossa
* *RGBA –* punainen, vihreä, sininen, alpha väritasokoodaus
* *Koko –* Elementin säteen pituus
* *Muoto –* Kuutio, pyramidi, kolmio, sylinteri, pallo
* *Tekstuuri –* Kuva, joka voidaan piirtää elementin pintaan

Vaikka iVIZ-järjestelmän arvioinnissa järjestelmän osoitettiin parantavan tiedon analysoinnin prosessia, järjestelmää ei silti voi kuvata vielä kovin innovatiiviseksi. Donalek ja kumppanit [2014] eivät tutkimuksessaan avanneet tai perustelleet tekemiänsä valintoja järjestelmän visualisointipäätöksiin liittyen. Tutkimuksen pohjalta voidaan vain todeta, että iVIZ piirtää dataelementtejä näkyviin hyödyntäen pisteparvi-visualisointia (*Scatter plot)*, joka erityisesti ison datajoukon kohdalla tekee visualisoinnista sekavan. iVIZ-järjestelmästä tarjolla olevat kuvankaappaukset itsessään jo osoittavat, että pisteparvi-tekniikka toimii heikosti myös virtuaalitodellisuutta hyödynnettäessä (Kuva 6). Donalek ja kumppanit mainitsevat tutkimuksensa lopuksi julkaisevansa alustan tiedeyhteisön käyttöön. Vaikuttaa kuitenkin siltä, että järjestelmän kehitystyö on lopetettu.

Toinen tieteellistä tutkimusta varten luotu Big Datan virtuaalitodellisuuden visualisointialusta luotiin Moranin ja kumppaneiden [2015] tutkimustyössä. Heidän tutkimuksessaan oli tavoitteena visualisoida tietoa MIT-kampuksen alueella lähetettyjen Twitter viestien sisällön pohjalta. Tutkimusryhmä mallinsi virtuaalitodellisuuteen kampusalueen ja Twitter-viestien geolokaatio-metatietoa hyödyntämällä, he pystyivät sijoittamaan viestin lähetyspaikan osaksi 3D-maailman sijaintia. Heidän tutkimuksessaan visualisoinnille ei asetettu muita määritteitä kuin data-elementin sijainti luodussa 3D-maailmassa pohjautuen mainittuun metatietoon. Täten visualisoinnin data elementit ovat vain kasattu päällekkäin niiden jakaessa saman geolokaation 3D-mallinnuksen sisällä (Kuva 10.). Moranin ja kumppaneiden tutkimuksellisena ongelmana on se, että he eivät varsinaisesti työstäneet Big Dataa tutkimuksessaan. Järjestelmässä kuvattiin vain Twitter-viestien sisältöä, joiden tietomalli on hyvin tunnettu, minkä myös heidän tutkimus vahvistaa: vain 2 % viesteistä sisälsi puutteita ja nämä jätettiin tutkimuksen ulkopuolelle. Tämän lisäksi tutkimuksessa käytetyn datajoukon määrä oli vain 6000 Twitter-viestiä ja data ei ollut reaaliaikaista. Näistä syistä myöskään heidän tutkimuksensa ei vastaa Big Datan visualisointiin liittyviin peruskysymyksiin eikä tutkimuksessa ole otettu kantaa näihin kysymyksiin miltään osin.



*Kuva 10. Moranin ja kumppaneiden [2015] visualisointialustan näkymät.*

Vuonna 2015 järjestetyn Big Data VR haasteen voittajaryhmä *Masters of Pie* kehittivät oman Big datan visualisointiin pohjautuvan järjestelmänsä, joka osoittaa aikaisempia tutkimuksia paremmin visuaaliset ja vuorovaikutteiset mahdollisuudet, joita virtuaalitodellisuudessa voitaisiin hyödyntää. Ryhmä huomasi jo kehitysprosessin alussa, että pelkästään datan ripottelu virtuaalitodellisuuteen, kuten Donalekin ja kumppaneiden [2014] ja Moranin ja kumppaneiden [2015] tutkimuksissa, ei ole merkityksellistä. Virtuaalitodellisuuteen luodun toteutuksen täytyisikin jo itsessään helpottaa analyysin tekemistä. Ryhmä päätyikin visualisoinnissaan DNA-ketjuun pohjautuvaan ratkaisuun, joka kiertyy spiraalin muotoisesti käyttäjän ympärille ja tuo suoraan kaiken tiedon käyttäjän näkyville. Masters of Pien ratkaisussa käyttäjälle ei tarjota mahdollisuutta liikkua vapaasti visualisoinnissa vaan luotetaan visualisoinnin toimivuuteen itsessään sekä tehokkaisiin vuorovaikutuksellisiin työkaluihin. Perinteisten valintojen ja suodatusten lisäksi, heidän ratkaisussa tarjotaan mahdollisuus datan eri ulottuvuuksien linkityksiin, jonka avulla käyttäjällä on mahdollisuus pyrkiä löytämään eri klustereita data joukon sisältä annettujen määritysten pohjalta. Masters of Pien toteutuksen taustatekijöitä esimerkiksi käytetyn datan osalta ei ole avattu, mutta heidän selvitystensä pohjalta käy selväksi, että käytettävän datajoukon sisältö on ollut etukäteen selvillä ja ainakin joiltain osin visualisointia on räätälöity kyseisen datajoukon mukaiseksi.



*Kuva 11. Masters of Pien visualisointityökalun toteutus, jossa kuvattuna on eri arvojen linkittäminen osaksi visualisoinnin eri ulottuvuuksia kuten kehän korkeus tai sijainti.*

4.3 Yhteenveto aikaisempien järjestelmien pohjalta

Esimerkkien läpikäymisen jälkeen voidaan todeta, että Masters of Pie ryhmän visualisointiratkaisu vaikuttaa tehokkaimmalta ja he ovat ratkaisussaan käyttäneet omaa innovaatiotaan. Järjestelmässä on keskitytty hyvään visualisointiin ja vuorovaikutuksellisiin työkaluihin, kun taas muiden alustojen ratkaisu pohjautuu vain datan sijoittamiseen maailmaan ja vapaaseen liikkuvuuteen sen ympärillä. Vapaan liikkumisen periaatetta ei voida nähdä datan analysoinnissa välttämättä perusteltuna, sillä analysoinnin vaiheessa, jota kyseiset järjestelmät palvelevat, halutaan nähdä datajoukon kokonaiskuva eikä niinkään olla kiinnostuneita yksittäisten elementtien arvoista. Täten erityisesti pääklustereiden löytäminen on datan käsittelyn alkuvaiheessa yksi tärkeimmistä tekijöistä, mikä voidaan ainakin osittain nähdä toteutuvan Masters of Pien toteutuksessa järjestelmän tarjoamien työkalujen avulla. Vapaa liikkuminen virtuaalitodellisuudessa voidaan nähdä myös mahdollisuutena vaihtaa näkökulmaa tietojoukkoa tarkasteltaessa. Donalekin ja kumppaneiden [2014] sekä Moranin ja kumppaneiden [2015] tutkimuksissa data on kuitenkin sijoiteltuna maailmaan matriiseja hyödyntäen, jolloin kuvakulman vaihtamisen merkitystä on vaikea perustella, tietoalkioiden ollessa staattisesti kiinnittyneinä koordinaatistoon.

Yhtenevää näillä kaikilla esimerkeillä on siinä, että yhdessäkään toteutuksessa ei oteta täysin kantaa tämän tutkimuksen alkuvaiheessa esitettyihin Big Datan hyödyntämisen ongelmakohtiin: Data ei ole reaaliaikaista, käytetyn datan sisältö on hyvin tai ainakin joiltain osin tiedossa ja datamäärät ovat hyvinkin maltillisia. Näiden tekijöiden takia aikaisemmat visualisointialustat tarjoavat heikon pohjan tässä tutkimuksessa tehtävää toteutusta ajatellen sekä yleisesti jatkotutkimusta silmällä pitäen.

Näiden esimerkkien pohjalta voidaan kuitenkin varmentaa jo aikaisemmin esille tullut tieto siitä, että Big Datalle suunnatun geneerisen ja datan sisältöön mukautuvan virtuaali­todellisuus­järjestelmän luominen on hyvin hankala prosessi. Näiden näkemysten pohjalta ei voida ajatella, että järjestelmä osaisi algoritmien pohjalta muodostaa aina oikean visualisoinnin annetulle datasyötteelle. Käyttäjän toiminnan ja tiedon välisen vuorovaikutuksen (suodatus, attribuuttien linkitykset) tulee olla pääpainotettuna myös visualisoinnin luomisessa, jos visualisoinnin prosessi halutaan mahdollistaa riippumatta datan sisällöstä. Tässä ajatuksessa palataan jälleen näkemykseen, että ihminen omaa erinomaisen kyvyn havaita datan sisäisiä rakenteita (*pattern*) ja riippuvuuksia näköaistinsa avulla, mikä koneellisesti vaatisi syvää prosessointia. Tällöin käyttäjälle annetaan vapaus poistaa häiriöt ja ottaa tarkasteluun vain häntä kiinnostavat osuudet, minkä turvin osa big datan visualisointia koskevista ongelmista saadaan eliminoitua. Edellä mainitut tekijät tullaan huomioimaan tämän tutkimuksen järjestelmän kehityksessä, joka käydään tarkemmin läpi kappaleessa 5. ja käydään läpi kompromissit sekä niiden perustelut, mitä mahdollisesti joudutaan kehityksen aikana tekemään eri riippuvuustekijöistä johtuen.

4.4 HTC Vive

Vive on HTC:n kehittämä virtuaalitodellisuuslasien teknologia. Järjestelmä tarjoaa tällä hetkellä kokonaisvaltaisemman virtuaalitodellisuusratkaisun kuin esimerkiksi kilpailija Oculus Rift. HTC Vive tarjoaa virtuaalitodellisuusnäkymän lisäksi käyttäjälle mahdollisuuden siirtää oma liikehdintä reaalimaailmassa suoraksi liikkeeksi virtuaalimaailmassa, jäljentämällä käyttäjän liikkumista erikseen määritellyn alueen sisällä infrapunakameroiden avulla. Lisäksi järjestelmä pystyy seuraamaan päässä olevien lasien ja käsissä olevien ohjainten liikehdintää. Taulukkoon 2 on koottu vertailu markkinoilla olevien virtuaalilasien ominaisuuksista. Taulukon pohjalta voidaan todeta, että HTC Vive on vaihtoehdoista tällä hetkellä selvästi monipuolisin ja tarjoaa suoraan ratkaisun interaktioon virtuaalimaailman sisällä ilman ulkopuolisten laitteiden kytkemistä osaksi järjestelmää.



*Taulukko 2. Markkinoilla olevien virtuaalilasien ominaisuuksien vertailu. (*[*www.virtuaalimaailma.fi/virtuaalilasit/*](http://www.virtuaalimaailma.fi/virtuaalilasit/)*)*

Aikaisemmissa, virtuaalitodellisuuteen pohjautuvissa visualisointijärjestelmissä, on hyödynnetty Oculus Rift -virtuaalilasiteknologiaa. Oculus Rift-virtuaalilasien lisäksi Donalekin ja kumppaneiden [2014] tutkimuksessa hyödynnettiin *Vicon* liikkeenseurantajärjestelmää [Vicon], jonka avulla käyttäjän liikkeet siirrettiin osaksi virtuaalista liikettä. Valintojen ja syötteiden antamiseen kaikissa tutkimuksissa [Donalek et. al 2014, Moran et. al 2015, Masters of Pie, 2015] hyödynnettiin *Leap Motionin* liikkeiden ja eleiden tunnistusjärjestelmää [Leap Motion]. Kuten mainittu, näitä vuorovaikutuksellisia elementtejä ja teknologioita silmällä pitäen Vive tarjoaa suoraan jo itsessään vaaditut ominaisuudet. Lisäksi Viven ominaisuudet voidaan helposti ottaa käyttöön Unity3D-kehitysympäristöön suunnattujen kirjastojen avulla, jolloin alustan kehitysaikaa säästyy huomattavasti. Näistä syistä tämän tutkimuksen alusta tullaan kehittämään Viven teknologian pohjalle. HTC Viven käyttöä tukee myös Masters of Pie ryhmän artikkelissa tekemä toteamus, jossa he toteavat, että alustan käyttökokemus olisi parempi HTC Viven tekniikalla toteutettuna.

5. TESTIJÄRJESTELMÄN KEHITYS

Testijärjestelmän tavoitteena oli luoda datan visualisoinnin toteutus virtuaalitodellisuuden ympäristöön sekä tätä vastaavana 2D-toteutuksena työasemaympäristöön. Visualisointidemoja luotiin kaksi erilaista: Interaktiivinen visualisointi maapallosta, johon reaaliaikaisesti päivittyy Twitter-viestejä näkyville niiden geolokaatiotietoihin pohjautuen. Lisäksi kaksi perinteisempää visualisoinnin toteutusta pylväs – sekä graafidiagrammeista. Testijärjestelmän kokonaisuudet kehitettiin Unityn 3D – pelimoottorilla, joka tukee erityisen hyvin virtuaalitodellisuusmaailmojen luomista [Unity]. Näiden kahden kokonaisuuden tarkempi sisältö sekä toteutusvaihe kuvataan seuraavaksi.

**5.1 Twitter-viestien visualisoiminen maailmankartalla**

Toiseksi esimerkiksi haluttiin luoda staattisemmasta visualisoinnista poikkeava, interaktiivisempi toteutustapa. Demon tarkoituksena on luoda syvempi visualisoinnillinen kokemus, jossa pystytään paremmin hyödyntämään virtuaalitodellisuuden interaktiotapoja kuten pään liikkeenseuranta. Visualisoinnissa haluttiin hyödyntää laajaa tietomallia, joka päivittyisi dynaamisesti käytön aikana. Lisäksi tietosisällön tuli olla monimuotoista, jotta visualisoinnilla pystyttäisiin tukemaan datan sisällön monimuotoisempaa esittämistapaa. Näiden kriteerien pohjalta visualisoitavaksi dataksi valikoitu Twitter-viestit, joita työssä hyödynnetään *Twitter Streaming API* – rajapinnan avulla [Twitter Stream API]*.* Twitter-viestien rajapinta tarjoaa mahdollisuuden suora toistaa ja kuunnella Twitteriin luotuja viestejä reaaliajassa (noin 6000 viestiä sekunnissa) ja tarjoaa täyden viesteihin liittyvän tietosisällön esimerkiksi viestin sisällön ja käyttäjän nimen. Näiden tietojen pohjalta viitaten Kappaleessa 2 esitettyyn Big Datan määrittelyyn, voidaan sanoa kriteerien toteutuvan Twitter Streaming API:n tuottavan datan kohdalla paitsi tiedon moninaisuuden (*Variety*) osalta. Rajapintoja hyödynnettäessä niistä saatava data ja sen rakenne on aina hyvin selkeästi kuvattu, jolloin käytettävän datan sisällön sekaisuus ei muodostu ongelmaksi ja se jätetään tässä tutkimuksessa huomiotta.

Käytettävään dataan ja tiedonhankintaan liittyvien päätösten tekemisen jälkeen seuraava vaihe oli miettiä, minkälaisen visualisaatiota sen ympärille voisi luoda. Twitter-viestien visualisoimista viestien geolokaatiotietojen pohjalta maailmankartalla on tehty jo aikaisemmin [Onemilliontweetmap], mutta tämä tarjosi hyvän lähtökohdan yksityiskohtaisemman visualisoinnin luomiselle, joka tarjoaisi syvällisemmän käyttökokemuksen myös virtuaalitodellisuuteen. Maapallon 3D–malli ja sen toiminnollisuudet otettiin käyttöön Unitylle luodusta World Political Map–kirjastosta [Kronnect, World Political Map]. Tämän jälkeen Twitter viestien käsittelyä varten luotiin oma rajapintansa ja viestien tiedot yhdistettiin osaksi maapallon 3D-mallia.



*Kuva 12. Havainnekuva Twitter-viestien visualisointidemosta.*

**5.1.1 Toiminnollisuus**

Visualisoinnissa tärkeimpänä tietona korostuvat Twitter–viestin sijaintikoordinaatit. Koordinaattien avulla viestit piirretään niiden lähettämispaikkaa vastaavaan sijaintiin maapallon 3D-mallissa. Käyttäjä voi tarkistella viestien sisältötietoja viemällä ohjattava kursori viesti-ikonin päälle. Kursorin ohjaaminen tapahtuu virtuaalitodellisuuden toteutuksessa virtuaalitodellisuuslasien liikeseurannan kautta ja työpöytätoteutuksessa hiiren avulla. Käyttäjä pystyy myös kääntämään maapalloa virtuaalitodellisuuden toteutuksessa erillisen ohjaimen avulla sekä työpöytäympäristössä hiiren painikkeen avulla. Virtuaalitodellisuuden ja työasemaympäristöjen eroavaisuudet on vielä kuvattu Taulukossa 3.

|  |  |  |
| --- | --- | --- |
| **Interaktio** | **Virtuaalitodellisuus** | **Työpöytäympäristö** |
| **Kursorin liikuttaminen** | Toteutettu virtuaalitodellisuuslasien liikkeenseurannan avulla. Kursori liikkuu maapallon pinnalla päätä liikuteltaessa. | Kursorin liikuttaminen hiiren avulla. |
| **Maapallon kääntäminen** | Toteutettu erillisen ohjaimen avulla. Pitämällä ohjaimen painiketta pohjassa ja liikuttamalla ohjainta, maapallo kääntyy ohjaimen liikettä vastaavaan suuntaan. | Pitämällä hiiren painiketta pohjassa ja kääntämällä maapalloa haluttuun suuntaan. |

*Taulukko 3. Virtuaalitodellisuuden ja työpöytäympäristön eroavaisuudet maapallovisualisoinnin demossa.*

**5.2 Twitter –viestitietojen visualisoiminen pylväs- ja graafidiagrammeilla**

Kehitettiin maapalloa kuvaava visualisointi, johon big datana ajetaan twitterin streaming api:n kautta tuotettua tietoa (pystytään helposti kontrolloimaan käytettävän datan määrää). Tarkoituksena on kehittää visualisoinnista virtuaalitodellisuudessa käytettävä versio sekä versio web serverilla ajettavaan 2D-työpöytäympäristöön.

- Tarkoituksena vertailla käyttäjäkokemusta virtuaalitodellisuuden ja 2D-ympäristön välillä.

- Mahdollisesti toteutettava myös jokin muu yksinkertaisempi visualisointi.

- Mitä tehtiin. Big dataa visualisoiva VR järjestelmä. Valmis data setti joka visualisoidaan, ei dynaaminen

- Miten tehtiin: Toteutettiin Unitylla, HTC Vive pohjaisena järjestelmänä. Kaksi eri tapaa visualisoida data setti ja käyttäjällä mahdollisuus kevyesti vuorovaikuttaa datan kanssa.

- Miksi tehtiin: Tarkoituksena kerätä käyttäjiltä kokemuksia VR:n data visualisointikokemuksista ja tyyleistä. Mikä olisi oikeanlainen tapa/lähtökohta lähteä visualisoimaan isoja datasettejä VR:ssä. Big Datan ongelmat, huomioitiinko tässä ja miten.

- Vaiheet: Määrittely mitä pitää tehdä, datan visualisointidesign, toteutus Unitylla

- Lopputulos: Mitä tuli tehtyä, kehtaako tätä edes testata käyttäjillä

6. TESTIJÄRJESTELMÄN KÄYTTÄJÄTESTAUS

- Testisuunnitelma, miten virtuaalitodellisuuden ja 2D-ympäristöjen eroja tässä testataan.

- Tulosten kertaaminen kahden eri visualisoinnin osalta.

- Haastattelukysymykset

- Yhteenveto

7. YHTEENVETO

7.1 Käyttäjätestauksen tulokset

- Mtä testauksen avulla kävi ilmi.

- Minkälaisen visualisointikokemuksen virtuaalitodellisuus tarjosi.

- Taulukoituja tuloksia, tehokkuudesta tai muusta vai pelkästään kvalitatiivinen testaus?

7.2 Loppuyhteenveteo

-Loppukaneetteja hahmoteltuna:

* Vaikka Big Datan visualisointi virtuaalitodellisuudessa kuulostaisi käytännössä hienolta ja tehokkaalta idealta, niin sitä se ei ole. Big Datan asettamat haasteet ovat itsessään jo haasteellisia visualisoinnille ja haasteet yhä kasvavat lisättäessä tekijöihin virtuaalitodellisuuden. Tämän lisäksi virtuaalitodellisuus ei kuitenkaan tuo mitään täysin uutta ja merkittävää pöytään (tulisiko keskittyä AR:n mahdollisuuksiin). Virtuaalitodellisuuden hyödyt voidaankin löytää lähinnä soveltuvan ja spesifin visualisoinnin käytöstä virtuaalitodellisuuden sisällä. Tällöin tämä kuitenkin tarkoittaisi sitä, että visualisointi tulisi luoda erikseen jokaiselle datajoukolle ja tällöin data joukon sisältö tulisi jo jollakin tapaa tuntea, jotta tiedon pohjalta voitaisiin luoda tehokas visualisointi VR maailman sisälle. Tällöin tavoitteet ja ratkaisut sulkevat toisensa pois. Virtuaalitodellisuuden visualisointi hyödyttää silloin, kun visualisointi on tarkka, spesifisempi ja vähemmän geneerinen. Geneerisen visualisointialustan luominen virtuaalitodellisuuteen ei tuo tarpeeksi uutta pöytään, mitä ei voitaisi hyödyntää jo työpöytäkoneiden 2D-maailmassa ja hyödyt jäävät pieniksi.

Kirjalliset viitteet

[Akerkar, 2013] Rajendra Akerkar, Big Data Computing. Taylor & Francis Group; 2013.

[Beck, 2003] Michael Beck, Real-Time Visualization of big 3D City Models. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol XXXIV-5/W10, Switzerland, February 2003.

[Burdea & Coiffet, 2003] Grigore C. Burdea, Philippe Coiffet, Virtual Reality Technology. Wiley, New Jersey 2003.

[Brown et al., 1995] Judith R. Brown, Rae Earnshaw, Mikail Jern, John Vince, Visualization: using computer graphics to explore data and present information. John Wiley & Sons, Inc., New York 1995.

[Buyya et al., 2009] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg…, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. Future Generation Computer Systems. 25. 599-616. 10.1016/j.future.2008.12.001.

[Card et al., 1999] Staurt K. Card, Jock Mackinlay, Ben Shneiderman, Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann Publishers, 1999.

[Chambers et al, 1983] John Chambers, William Cleveland, Beat Kleiner, Paul Tukey, Graphical Methods for Data Analysis. Wadsworth, 1983.

[Chen & Liu, 2004] Keke Chen and Ling Liu. 2004. VISTA: validating and refining clusters via visualization. Information Visualization 3, 4, 257-270, December 2004.

[Chen & Zhang, 2014] C.L. Philip Chen, Chun-Yang Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, Information Sciences, 275 (10), pp. 314-347, August 2014.

[Chen, 2014] Keke Chen. 2014. Optimizing star-coordinate visualization models for effective interactive cluster exploration on big data. Intell. Data Anal. 18, 2,117-136, March 2014.

[Cox et al., 1997] Donna J. Cox, Robert M. Patterson, Marcus L. Thiebaux, Virtual reality 3D interface system for data creation, viewing and editing. Patentti: <https://www.google.ch/patents/US6154723>

[Cruz-Neira et al., 1993] Carolina Cruz-Neira, Daniel J. Sandin, Thomas A. DeFanti, Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. Proceedings of the 20th annual conference on Computer graphics and interactive techniques, 135-142, 1993.

[Donalek et al., 2014] Ciro Donalek, S.G. Djorgovski, Scott Davidoff, Alex Cioc ja muut. Proceedings of 2014 IEEE International Conference on Big Data, p 609-615, October 2014.

[Du et al., 2016] Fei Du, A-Xing Zhu, Feng Qi, Interactive visual cluster detection in large geospatial datasets based on dynamic density volume visualization. Geocarto International, 31:6, 597-611, 2016.

[Dutcher, 2014] Jennifer Dutcher, What is Big Data, 2014 Berkley School of Information. https://datascience.berkeley.edu/what-is-big-data.

[Faber, 2013] Brenton Faber, Analytics, Big Data, and Technical Communication. Esitetty: *Association of Teachers of Technical Writing Conference*. Las Vegas, NV. March 2013.

[Fernandez et al., 2014 ] Alberto Fernandez, Sara del Rio, Victoria Lopez, Abdullah Bawakid, Maria J. del Jesus… Big Data with Cloud Computing: an insight on the computing environment, MapReduce, and programming frameworks. WIREs Data Mining Knowl Discov, 4: 380–409, 2014.

[Few & Edge, 2007] Stephen Few, Perceptual Edge, Data visualization: past, present, and future. IBM Cognos Innovation Center, 2007.

[Fisher, 2016] Danyel Fisher, Big data exploration requires collaboration between visualization and data infrastructures. HILDA’16 Proceedings of the Workshop on Human-In-the-Loop Data Analytics. San Francisco, California, July 2016.

[Foo et al., 2009] Jung-Leng Foo, Thom Lobe, Eliot Winer, A Virtual Reality Environment for Patient Data Visualization and Endoscopic Surgical Planning. Journal of Laparoendoscopic & Advanced Surgical Techniques, vol 19, supplement 1, 211-218, 2009.

[Haber & McNabb, 1990] Robert B. Haber, David A. McNabb, Visualization Idioms: A Conceptual Model for Scientic Visualization Systems. In: G.M. Nielson,B. Shriver, L. Rosenblum (eds.) Visualization in Scientic Computing, IEEE Computer, 74{93. IEEE Computer Society Press Los Alamitos, CA, USA (1990)

[Hadi et al., 2014] Hiba Jasim Hadi, Ammar Hameed Shnain, Sarah Hadishaheed, Azizahbt Haji Ahmad, Big Data and Five V’s Characteristics. Big Data and Five V’s Characteristics. Proceedings of IRF International Conference, 1st November 2014, Tirupati, India.

[Harrison & Römhild, 2007] Chris Harrison, Christoph Römhild, Bible Cross-References. http://www.chrisharrison.net/index.php/Visualizations/BibleViz

[IBM, 2013] IBM: The Four V’s of Big Data, 2013. http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[InfoVis] InfoVis Wiki, <http://www.infovis-wiki.net/index.php?title=Visualization_Pipeline>

[Intel, 2013] Intel, Big Data Visualization: Turning Big Data Into Big Insights. https://www.intel.com/content/dam/www/public/us/en/documents/white-papers/big-data-visualization-turning-big-data-into-big-insights.pdf

[International Data Corporation, 2014] International Data Corporation, Data Growth Business Opportunities, and the IT Imperatives. April 2014, https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.html

[Johnson et al., 2005] Chris Johnson, Robert Moorhead, Tamara Munzner, Hanspeter Pfister, Penny Rheingans, Terry S. Yoo. 2005. NIH-NSF Visualization Research Challenges Report. Los Alamitos, California: IEEE.

[Kandogan, 2000] Eser Kandogan, Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions. Proceedings of the IEEE information visualization symposium, late breaking hot topics, p. 9–12, 2000.

[Kandogan, 2001] Eser Kandogan, Visualizing multi-dimensional clusters, trends, and outliers using star coordinates, Proceedings of ACM SIGKDD Conference, 107–116, 2001.

[Keahey, 2013] T. Alan Keahey, Using visualization to understand big data. Technical Report, IBM Corporation, 2013, 1-16.

[Kehrer et al., 2012] J. Kehrer, R.N. Boubel, P. Filzmoser, H. Piringer, A Generic model for the integration of interactive visualization and statistical computing using R. Conference on Visual Analytics Science and Technology (VAST), IEEE 2012 p. 233-234

[Khan & Khan, 2011] Muzammil Khan, Sarwar Shah Khan, Data and Information Visualization Methods and Interactive Mechanisms: A Survey, International Journal of Computer Applications, 34(1), 1-14, 2014.

[Kim, 2005] Gerard Jounghyun Kim, Designing Virtual Reality Systems, The Structured Approach. Springer London, London 2005.

[Klein & Gorton, 2015] John Klein, Ian Gorton, Runtime Performance Challenges in Big Data Systems. Proceedings of the 2015 Workshop on Challenges in Performance Methods for Software Development. January 2015.

[Laney, 2001] Doug Laney, 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group, February 2001.

[Long & Linsen, 2011] Tran Van Long, Lars Linsen, Visualizing high density clusters in multidimensional data using optimized star coordinates. Computational Statistics 26:655, December 2011.

[Marcos et al., 2013] Marcos D. Assuncao, Rodrigo N. Calheiros, Silvia Bianchi, Marco A. S. Netto..., Big Data Computing and Clouds: Trends and Future Directions. Journal of Parallel and Distributed Computing, Volumes 79-80, 3-15, May 2015.

[Miksch & Aigner, 2014] Silvia Miksch, Wolfgang Aigner, A matter of time: applying a data-users-tasks design triangle to visual analytics of time-oriented data. Comp Graph, 2014, 38:2, 86-90.

[Monino, 2016] Jean-Louis Monino, Data Value, Big Data Analytics and Decision-Making. Journal of the Knowledge Economy, 1-12, 2016.

[Moran et al., 2015] Andrew Moran, Viljay Gadepally, Matthew Hubbell, Jeremy Kepner, Improving Big Data visual analytics with interactive virtual reality. IEEE High Performance Extreme Computer Conference (HPEC), pp. 1–6, 2015.

[Moreland, 2013] Kenneth Moreland, A Survey of Visualization Pipelines, IEE Transactions on visualizations and computer graphics, vol. 19, no. 3, March 2013.

[Muller & Schumann, 2003] Wolfgang MulIer, H. Schumann, Visualization method for time-dependent data—an overview. Proceedings of the 2003 Winter Simulation Conference, vol. 1. IEEE; 2003.

[Neves & Bernardino, 2015] Pedro Neves, Jorge Bernardino, Big Data Issues. IDEAS’15 Proceedings of the 19th International Database Engineering & Applications Symposium, p 200-201, Yokohama, Japan, July 2015.

[Ngrain, 2013] Ngrain, 3 reasons why “visualization” is the biggest “V” for big data. https://actuariatfinance.wordpress.com/2014/12/01/3-reasons-visualization-biggest-v-big-data/

[Office of Science and Technology Policy, 2012] Office of Science and Technology Policy 2012, Obama administration unveils "big data" initiative: Announces $200 million in new R&D investments. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big\_data\_press\_release\_final\_2.pdf

[Olshannikova et al., 2015] Ekaterina Olshannikova, Aleksandr Ometov, Yevgeni Koucheryavy, Thomas Olsson, Visualizing Big Data with Augmented and Virtual Reality: Challenges and Research Agenda. Journal of Big Data, Vol. 2, 22, 01.10.2015.

[Playfair, 1786] William Playfair, The Commercial and Political Atlas. London, 1786.

[Plugfelder & Helmut] Plugfelder, Ehren Helmut, Big Data, Big Questions. *Communication Design Quarterly* 1, no. 4, 18-22, 2013.

[Rajaraman, 2016] V. Reson Rajaraman, Big Data Analytics. Resonance, August 2016, Volume 21, Issue 8, 695 – 716.

[Rijmenami, 2016] Mark van Rijmenam, Why The 3V’s Are Not Sufficient To Describe Big Data. https://datafloq.com/read/3vs-sufficient-describe-big-data/166. 2016

[Rubio-Sanchez & Sanchez, 2014] Manuel Rubio-Sanchez, and Alberto Sanchez, Axis Calibration for Improving Data Attribute Estimation in Star Coordinates Plots. IEEE Transactions on Visualization and Computer Graphics, vol. 20, no. 12, pp. 2013-2022, December 31 2014.

[Shneiderman, 1992] Ben Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach. Journal ACM Transactions on Graphics, 11,1, 92-99, January 1992.

[Shvachko et al., 2010] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler, The Hadoop Distributed File System. MSST '10 Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST).

[Sucharitha et al., 2014] Venkataramana Sucharitha, S.R. Subash, P. Prakash, Visualization of Big Data: Its Tools and Challenges, International Journal of Applied Engineering Research, 9(18), pp. 5277-5290, 2014.

[Sutherland, 1968] Ivan E. Sutherland, A Head-mounted three dimensional display. Proceedings of the AFIPS Fall Joint Computer Conference, Thompson Books, 757-764, Washington D.C, 1968.

[Szewczyk et al., 2004] Rovert Szewczyk, Joseph Polastre, Alan Mainwaring, David Culler, Lessons from a sensor network expedition. European Workshop on Wireless Sensor Networks, 307-322. 2004.

[Teng et al., 2015] Pengju Teng, Hongjun Li, Xiaopeng Zhang, Survey on Visualization Layout for Big Data. International Conference on Intelligent Science and Big Data Engineering, [Intelligence Science and Big Data Engineering. Big Data and Machine Learning Techniques](https://link.springer.com/book/10.1007/978-3-319-23862-3) pp 384-394, October 2015.

[Teräs & Raghunathan, 2015] Marko Teräs, Shriram Raghunathan, Big Data Visualisation in Immersive Virtual Reality Environments: Embodied Phenomenological Perspectives to Interaction. International Journal of Soft Computing. 05. 1009-1015, August 2015.

[Tolle et al., 2005] Gilman Tolle, Joseph Polastre, Robert Szewczyk ja muut, A macroscope in the redwoods. SenSys ’05 Proceedings of the 3rd international conference on Embedded Networked Sensor Systems, 51-63. November 2005, San Diego, California, USA.

[Turner et al., 2014] Vernon Turner, J. F. Gantz, D. Reinsel, S. Minton, The digital universe of opportunities: rich data and the increasing value of the internet of things. IDC EMC, 2014. https://www.emc.com/leadership/digital-universe/2014iview/digital-universe-of-opportunities-vernon-turner.htm

[Wang et al., 2015] ]Lidong Wang, Guanghui Wang, Cheryl Ann Alexander, Big Data and Visualization: Methods, Challenges and Technology Progress. Digital Technologies. Vol. 1, No. 1, 2 33-38, 2015.

[Ware, 2004] Colin Ware, Information Visualization – Perception for Design. Morgan Kaufmann, April 2004.

[Watson, 2014] Hugh J. Watson, [Tutorial: Big Data Analytics: Concepts, Technologies, and Applications](http://elibrary.aisnet.org/Default.aspx?url=http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3785&context=cais). Communications of the Association for Information Systems: Vol. 34 , Article 65, 2014.

[Vo et al., 2011] Huy T. Vo, Jonathan Bronson, Brian Summa, Joao L.D Comba..., Parallel visualization on large clusters using MapReduce. Large Data Analysis and Visualization (LDAV), 2011 IEEE Symposium on, Providence, Rl, 2011, 81-88.

[Yahoo, 2016] Yahoo Confirms at Least 500 Million Accounts Were Hacked. Fortune.com/2016/09/22/yahoo-hack

**Sähköiset viitteet**

Vicon, Motion Capture Systems. <https://www.vicon.com/>

Leap Motion. <https://www.leapmotion.com/>

Twitter Stream API. <https://developer.twitter.com/en/docs>

Unity 3D. <https://unity3d.com/>

Kronnect, World Political Map – Globe Edition. Editor Extension. [https://www.assetstore.unity3d.com/en/#!/content/41890](https://www.assetstore.unity3d.com/en/%23!/content/41890)