Riku Hiltunen

Liiketoimintatiedon hallintaratkaisut ja yrityksen tietovarastojen hyötykäytön yleisimmät ongelmat

Tietotekniikan kandidaatintutkielma

1. helmikuuta 2019

Jyväskylän yliopisto

Informaatioteknologian tiedekunta

Tekijä: Riku Hiltunen

Yhteystiedot: riku.hiltunen@live.com

Ohjaaja: Sanna Mönkölä

Työn nimi: Liiketoimintatiedon hallintaratkaisut ja yrityksen tietovarastojen hyötykäytön yleisimmät ongelmat

Title in English: Business Intelligence and common problems with corporate data warehouses

Työ: Kandidaatintutkielma

Sivumäärä: 28+0

Tiivistelmä: Kandidaatintutkielmassani käyn kirjallisuuskatsauksen muodossa läpi, mitä ongelmia data-analyytikon tulee huomioida liiketoimintatiedon hallintaratkaisua toteutettaessa. Tavoitteenani on tutkia, minkä tyyppistä taustatyötä yrityksen tietokantojen yhdistämiseen tavallisesti kuuluu, jotta siitä voidaan tuottaa järkeviä otejoukkoja tiedonlouhintaan. Tutkin alan kirjallisuudessa normeiksi todettuja toimintamalleja esimerkkien avulla. Tarkastelun kohteena on esimerkiksi virheellisten ja puuttuvien arvojen käsittely otejoukossa.

Avainsanat: Tietovarastointi, massadata, visualisointi, likainen data, liiketoimintatiedon hallinta, ETL

Abstract: In this bachelor's thesis, I research what needs to be taken in consideration according to literature, when implementing a business intelligence solution. I approach the problem from data-analyst's perspective. My goal is to find out, what kind of work is included on combining company's databases. This needs to be done, in order to make a sensible data set for data mining, which business intelligence is. In this thesis, I present approved methods from the field's literature, through examples.

Keywords: Data storage, big data, visualization, dirty data, business intelligence,

ETL

Kuviot

Kuvio 1. Tietokannan ongelmia yksikkötasolla. Rahm & Hai Do (2000)..... 7
Kuvio 2. Tietokannan ongelmia skeematasolla (eli rivi/taulu) Rahm&Hai Do (2000)..... 8
Kuvio 3. ETL - kaavio tietovarastosta. Rahm&Hai Do (2000).....15

Sisältö

1 JOHDANTO LIIKETOIMINTATIEDON HALLINTAAN 1

2 LIIKETOIMINTATIEDON HALLINNAN TARJOAMAT MAHDOLLISUUDET 3
2.1 Liiketoimintatiedon hallintaratkaisun toteuttamisen haasteet 4
2.2 Massadatan haasteet 4

3 TIETOVARASTOJEN KÄYTTÖ 6
3.1 Yksikkö-/solutason ongelmia 7
3.2 Skeematason ongelmia 8
3.3 Tietokantojen yhdistämisen ongelmia tietovarastossa 9
3.4 Tyhjien arvojen käsittely ja tutkimuskysymysten asetus 10
3.5 Virheelliset ja puuttuvat arvot 11

4 RATKAISUMALLIT 14
4.1 ETL-prosessin askeleet likaisen datan poistamisessa 14
4.1.1 Nouto 14
4.1.2 Muunnos 15
4.1.3 Lataus 16
4.1.4 Tietovarastointi 16
4.2 Ulkoasu ja visuaalinen suunnittelu 16
4.3 Ennusteet 17

5 YHTEENVETO 19

KIRJALLISUUTTA 20

1 Johdanto liiketoimintatiedon hallintaan

Liiketoimintatiedon hallinta (BI, engl. *Business Intelligence*), usein sisältää visualisoidun raportointityökalun (dashboard, engl. *BI-dashboard*), jolla voidaan louhia yrityksen omistamasta datamassasta tietoa nopeasti ja tehokkaasti. BI-työkalu käsittelee tietokannoista noutamansa datan yhtenevään muotoon, jonka se esittää visuaalisessa muodossa. Tuolloin yrityksen datamassasta saadaan tietoa, jonka perusteella kyetään toimimaan ja tekemään päätöksiä. Tuotetusta datasta voidaan havaita siihen piiloutunutta tietoa ja riippuvuussuhteita, joiden perusteella yrityksen toimintaa voidaan kehittää.

Liiketoimintatiedon hallinta voidaan siis nähdä yrityksen tuottaman datan hyötykäyttönä. BI:n avulla saadaan lisäarvoa yrityksen toimintaan (Hokkanen 2012), kun yhtiön senhetkinen tilanne ja tarina nähdään visualisoituna ymmärrettävään muotoon. BI-ratkaisujen käyttöliittymiin päästään tyypillisesti verkkoselaimella, tai mobiililaitteilla. Raportointinäkymät ja sovellus tulisi tuottaa siten, että käyttäjä voi ohjata sovellusta näyttämään haluamaansa tietoa. Sovellus koostuu sivuista, joilla on elementtejä, kuten piirakkakaavioita. Piirakkakaaviosta valittaessa tarkastelun kohde, päivittyvät muut sivun elementit näyttämään kyseisen kohteen kuluja, kehitystä ja muita visualisointeja.

Havainnollistetaan BI:n tarjoamia mahdollisuuksia esimerkiksi huoltoyhtiön avulla. Huoltoyhtiön johtaja haluaa tietää, kuinka yhtiö on toiminut ja miten sen tulos on kehittynyt. BI-työkalujen avulla voidaan tähän kysymykseen vastata reaaliajassa, yhdistämällä huoltoyhtiön tietokannat BI-alustaan. Oli kyse huoltotehtävistä, henkilöstökuluista, tai tuottavuusindekseistä, voidaan BI:n avulla koota reaaliaikainen raportti yrityksen tilasta, sen tuottaman datan pohjalta. Luotua raporttia voidaan tarkastella eri näkökulmista. Esimerkiksi alueittain, vasteajoittain, tai vaikka kuukausittaisen tuottavuuden kehityksen perusteella. Datan pohjalta voidaan jopa luoda ennusteita tulevaisuuteen, esimerkiksi miten sääennusteiden voidaan olettaa vaikuttavan tarvittavaan työvoiman ja kaluston määrään. Yhdistämällä yrityksen tietokantoihin ilmatieteen laitoksen avoimet säätietokannat, voidaan BI-raportissa

luoda ennusteita tarvittavalle työvoimalle, tai priorisoida huoltotöitä venyneiden vasteaikojen perusteella. Yleisesti tunnettuja BI-työkaluja ovat Microsoft PowerBI, Qlik, Tableau ja IBM Cognos.

2 Liiketoimintatiedon hallinnan tarjoamat mahdollisuudet

Liiketoimintatiedon hallinta (BI, engl. *business intelligence*) on systemaattista organisaation ulkoisiin ja sisäisiin tietovarantoihin kohdistuvaa toimintaa (Korhonen 2012). Tällä tarkoitetaan pääsääntöisesti tietokantojen analysointia ja käsittelyä tulkittavampaan ja ymmärrettävään muotoon. Yrityksen sisäisten tietokantojen BI-ratkaisuilla voidaan tarjota yrityksen johdolle kykyä tehdä päätöksiä, perustuen asianmukaisiin pohjatietoihin(Hansoti 2010). Kolmansien osapuolien tietokantojen avulla voidaan luoda tuotetulle palvelulle lisäarvoa, yhdistämällä useamman palveluntarjoajan ja asiakkaan dataa. Toiminnan tehostamisen ohessa tuotetetaan myös yrityksen maineelle lisäarvoa ulkoisten toimijoiden näkökulmasta (*CRM Customer relationship management*) (Korhonen 2012).

Dresner Advisory Services (2017) tuottaman tutkimuksen tuloksissa huomattiin, että yrityksien tietovarastojen hyötykäytön ja BI:n käyttöönoton kasvu on ollut kovaa. Columbus2017 analysoi, että samassa yrityskohderyhmässä(US) vuonna 2015, 17 prosenttia yrityksistä hyödynsi BI:tä, mutta vuonna 2017, 53 prosenttia vastanneista yrityksistä hyödynsi BI:n tarjoamia mahdollisuuksia. Columbus 2017 toteaa, että päätavoitteet BI:n käyttöönotossa ovat olleet raportointi, raportointinäkymät, näiden edistyneet visualisoinnit, loppukäyttäjälle tarjottava itsepalvelu ja tietovarastointi. BI-työkaluilla tuotettu raportti, tai raportointinäkymä on interaktiivinen. Raportointinäkymä hyödyntää usein päivittäin noudettua ja käsiteltyä dataa yrityksen tilasta. Tällöin yrityksen johdolla on tuore tilannekuva yrityksensä toiminnasta, eikä päätöksiä tehdä vanhan tiedon perusteella. (Chen, Roger, Chiang ja Veda 2012) Näin saavutetaan kilpailullista etua, parantamalla suhdetta asiakkaisiin tukemalla heidän päätöksentekoaan (Korhonen 2012). BI-ratkaisulla voidaan tukea myös oman yrityksen johdon päätöksentekoa (Chen ym. 2012)(Hokkanen 2012).

2.1 Liiketoimintatiedon hallintaratkaisun toteuttamisen haasteet

Sahayn (2017) mukaan tiedon visualisointi on olennainen osa datan analysointia. Muita yleisiä keinoja hänen mukaansa data-analytiikassa ovat koneoppiminen, tilastotieteelliset testit, sekä tiedonlouhinta. Näistä hän luokittelee datan visualisoinnin tehokkaimmaksi ja helpoimmaksi tavaksi käsittää suuren otoksen sisältämää tietoa. Visualisoidusta datasta on helppo siirtyä seuraavaan vaiheeseen tietovaraston analysoinnissa, kun huomiota herättävät kohteet on ensin tunnistettu ja näistä haluttavat kysymykset on asetettu. Tyypillisesti saadut tulokset integroidaan BI-ratkaisuun.

Data-analyytikon vastuulla on päättää, miten tietokannasta havaittuja ongelmia ja johtopäätöksiä käsitellään (Osborne (2013); Sahay (2017)). Visualisoinnilla voidaan esimerkiksi louhia tehokkaasti datasta löytyviä tietoja, kaavoja, tai riippuvuussuhteita (Sahay 2017). Tuolloin tulee pohtia, kuinka löydetty tieto tuodaan esille ja mitä johtopäätöksiä siitä voidaan vetää (Sherman 2014). Aiheesta voidaan tehdä näiden johtopäätösten perusteella jatkotutkimusta.

Jatkotutkimus ja tiedon esilletuonti voidaan hoitaa pätevällä BI-työkalulla. Koneoppimisen avulla voidaan tutkia visualisointiprosessissa löytynyttä tietoa, samasta otejoukosta, jonka tuloksista saadaan uusi otejoukko käsiteltäväksi(Sahay 2017). Koneoppimisalgoritmeilla voidaan luoda ennusteita käsiteltävän datan pohjalta, jotka voidaan usein integroida BI-työkalun raportointinäkymään. Riippuvuussuhteita ja yleisimpiä koneoppimisella otejoukosta saatavia tietoja, on helppo löytää graafisen työkalun avulla, kuten RapidMiner. BI-raporttiin voidaan nämä löydetyt louhinnan kohdealueille tehtävät toimenpiteet sisällyttää. Data-analytiikan kaikilla osaalueilla(visualisointi, tiedonlouhinta, koneoppiminen) otetaan samat alkuaskeleet, sillä yleisin ongelma tietokannoissa, on likainen data (Osborne 2013).

2.2 Massadatan haasteet

Yrityksen menestymisen ja tuottavuuden kulmakivenä on yhä kasvavissa määrin, niiden tuottaman ja vastaanottaman datan hyötykäyttö. Omistettua dataa osataan

harvoin käsitellä tehokkaasti, sillä siivoamattomana ja analysoimattomana on sen hyödyntäminen hankalaa (Sherman 2014). Yrityksen tietovarastot ovat usein monesta lähteestä ja pitkältä aikaväliltä. Dataa voi olla yrityksen itse tuottamana ja esimerkiksi eri palveluntarjoajilta ja asiakkailta. Vuosien välissä merkintätavat ja yksiköt ovat saattaneet muuttua asiakkailla, palveluntarjoajilla, tai yrityksellä. Tietovaraston koko on tällöin suuri, monessa muodossa ja varmasti epäyhteneväinen. Sherman (2014) toteaa, että raakadatalla ei itsessään ole mitään arvoa, mutta siivotulla ja louhitulla datalla on. Tästä voidaan johtaa tietoa, jonka perusteella voidaan toimia. Käyttäjien syötteestä tuotettu, tai useammasta lähteestä yhdistetty raakadata on käytännössä aina likaista dataa (Sherman 2014).

3 Tietovarastojen käyttö

Likaista dataa (engl.*Dirty data*) (Sherman 2014)) syntyy jo yksittäisissä tiedostoissa, tai tietokannoissa. Kuviossa 1 (Rahm&Hai Do 2000) esitellään yksikkötasolla mahdollisesti esiintyvää likaista dataa tietokannassa ja kuviossa 2 esitellään likaista dataa skeematasolla. Visuaalista BI-ratkaisua tehdessä, data halutaan puhdistaa ja luoda tästä interaktiivinen raportointinäkymä.

Jotta tietovaraston sisältämien tietokantojen sisällöstä voidaan vetää johtopäätöksiä (Rahm&Hai Do 2000), on sen puhtauden oltava päätavoite. Muutoin tietovarastosta louhittava tieto saattaa olla virheellistä (Rahm&Hai Do 2000). Heterogeeniseen tietovarastoon siirryttäessä, puhdistustarpeet käsiteltävässä datassa kasvavat moninkertaisiksi(Hernández& Stolfo 1998). Hernándezin (1998) mukaan yhtenä yleisimpänä esimerkkinä toimivat otteessa ilmentyvät toisteisuudet (Kuvio 1), johon Rahm ja Hai Do (2000) lisäävät puuttuvan (engl.*missing information*) datan.

Yrityskaupan jälkeen ostajayrityksen tulee yhdistää ostetun yrityksen tietokannat omiin tietokantoihinsa. Ostetun yrityksen ollessa esimerkiksi paikallinen toimija, on sille riittänyt kirjata toiminnassaan vain kadunnimet tai postinumerot. Kansainvälisessä ostajayrityksessä tapana on ollut merkitä toimintaan mukaan myös maa. Herää kysymys, montako kauppakatua löytyy Suomesta. Tai montako kirkkokatua löytyy Keski-Suomesta. Vastaavat esiintymät tietokannassa synnyttävät toisteisuutta, joista jokaiselle esiintymälle tulisi löytää uniikki avain (Osborne 2013).

Massadata tulee identifioida (Osborne 2013), jotta BI-ratkaisu osaa huomioida toisteisina ilmenevien esiintymien erot. Ostetun yrityksen data tulee siivota ennen integraatiota. BI-ratkaisusta saattaa puuttua ostetun yrityksen data kokonaan, jos otejoukolle asetetut eheysrajoitukset (esim. tyhjät kentät) siihen sisällytettävästä datasta eivät täyty.

Yrityksen johto haluaa tietää, miten yrityksen kuljetuskaluston sijoittelua voidaan optimoida. Edellämainitussa ongelmatilanteessa kuljetuskalusto ja tilaukset näkyisivät todennäköisesti väärissä sijainneissa, tai eivät ollenkaan virheellisen da-

Ongelma	Likainen data	Syyt
Puuttuvat arvot	puh= "999-999", *puh=null*	Virheellinen arvo, tai kenttä tyhjänä
Kirjoitusvirheet	kaupunki="Jyvääskylä	Kirjoitusvirhe, esimerkiksi murteen takia
Lyhenteet ja epäselvät kuvaukset	kokemus="B"; työ="A Ma."	
Useammat arvot samassa solussa	nimi="Ville V. 12.01.98 Tampere"	Vapaan syötteen kenttä
Syöte väärässä solussa	kaupunki="Suomi"	Tieto väärässä kentässä
Riippuvuusvirheet solujen välillä	kaupunki="Jyväskylä", posNro="00100"	Postinumeron tulisi vastata kaupunkia
Epäyhtenevät merkintätavat	nimi="V. Virtanen", nimi²="Riku R."	Tyypillinen virhe, syntyy varsinkin vapaan syötteen kentissä
Toisteisuus kirjoitusvirheen takia	hlö=[nimi="V. Virtanen", ..] hlö²=[nimi="Ville Virtanen", ..]	Samasta henkilöstä on kaksi esiintymää, eri arvoilla (kirjoitusvirhe)
Toisteisuus	hlö=[nimi="Vili Virta", sPai="12.12.92"] hlö²=[nimi="Vili Virta", sPai="12.02.92"]	Samasta henkilöstä on kaksi esiintymää, eri arvoilla (kumpi oikea?)
Väärät referenssit	hlö=[nimi="Vili Virta", osasto="17"	Viitattu osasto on väärä

Kuvio 1. Tietokannan ongelmia yksikkötasolla. Rahm&Hai Do (2000)

tan vuoksi (Rahm&Hai Do 2000). Ostetun yrityksen data tulee käsitellä, jotta data olisi yhtenäistä yritysten välillä. Jos datan formaattiin, tai sen formatointiin ei puututa, tietovarannosta saatu tieto tulee näyttämään virheelliseltä yhdistetyn otejoukon osajoukkojen ollessa eri muodoissa (Osborne 2013).

Yhtiön arkipäiväisessä toiminnassa likaisen datan syntyminen on nykyisin normi. Sen syntymistä voidaan pyrkiä hallitsemaan rajoituksilla, mitä tietokantaan dataa syöttäviltä ohjelmilta vaaditaan. Huolimattomuuteen ja yksikkötason virheisiin tulee puuttua muilla keinoin (Kuvio 1). Puhtaan datan avulla voidaan tukea yritysjohdon laskentatoimea (Hokkanen 2012). Rahm ja Hai Do(2000) nimittävät likaisen datan tietokantoja ”garbage in, garbage out” - tyyppisiksi tietokannoiksi.

3.1 Yksikkö-/solutason ongelmia

Tietokantoja yhdistäessä, dataa yhdistetään tyypillisesti avaimien avulla. Avaimet luodaan attribuuttien ja arvojen yhdisteistä, joille pyritään löytämään vastaavuudet tietokantojen välillä. Ongelmat liittyen vanhan datan yksikkötason ongelmiin silti säilyvät. Yrityksen osastoilla on saattanut olla eriävät merkintätavat, tai asiakkaan tilin vaihtuessa on luotu vahingossa rinnalle uusi asiakas. Uudelle asiakkaalle on työntekijöiden ”vanhasta muistista” kirjattu toimintaa vanhan tunnisteen rinnalle. Tuolloin syntyy toisteisia esiintymiä (Hernández ym. 1998), joiden löytäminen kan-

Ongelma	Likainen data	Syyt
Laiton arvo	syntPaiva=30.13.1970	Arvot rajoitusten ulkopuolella (kk !=13)
Laiton arvo, riippuvuuden rike	ika=22, syntPaiva=30.11.1971	ika = (tamaPaiva-syntPaiva) tulee olla tosi
Yksilöivän tunnisteen rike	hlö=["Ville Virtanen", ID="12345"] hlö=["Riku Riemukas", ID="12345"]	ID:n tulee olla uniikki.
Viitteen rike	hlö=["Ville Virtanen", osasto="93"]	Osastoa 93 ei löydy.

Esimerkkejä tietokannan eheysongelmista

Kuvio 2. Tietokannan ongelmia skeematasolla (eli rivi/taulu) Rahm&Hai Do (2000)

nasta saattaa olla haastavaa. Tuolloin aiemman esimerkin kuljetuskalusto saatetaan siirtää vääriin sijainteihin, tai yritykselle syntyy haamukalustoa, joka on olemassa vain tietokantavirheissä. Vaihtoehtoisesti kirjatun toiminnan sijainniksi on saatettu merkitä [”Atmpere”/ ”Tampere”/ ”Tre”]. Näin ollen liiketoimintaa jää ulos haussa, jossa haetaan kaikki avaimen ”Tampere” - sijainnissa tapahtunut toiminta.

3.2 Skeematason ongelmia

Yksittäisiä tietokantoja yhdistettäessä tietovarastoksi, mahdollisuudet likaisen datan tuomiselle ja syntymiselle varastossa, ovat suuret (Rahm&Hai Do 2000). Tietovarastot jatkuvasti päivittävät ja hakevat suuria määriä dataa useasta lähteestä. Vosburg& Kumar (2004) artikkelin mukaan tyypillisesti likaista dataa syntyy, kun yrityksen osastot käyttävät omia tietokantojaan, joissa tiedon merkintätavat eivät ole yhteneväisiä. Ajatellaan tilanne, jossa samalle asiakkaalle on merkitty useampi tili, tai monta osoitetta. Jotkut osastot päiväävät tilaukset suomalaisessa ”DD/MM/YYYY” -formaatissa ja toiset osastot päiväävät ne yhdysvaltalaisessa ”MM/DD/YYYY” -formaatissa. Samalle asiakkaalle on saatettu myös käyttää eriäviä tunnistenumeroita osastojen välillä, jolloin jokaiselle asiakkaalle ei löydy samaa yksilöityä tunnistetta.

Tätä on tyypillisesti jatkunut yrityksessä vielä pitkään, jolloin syntynyt "likainen data"on tullut normiksi(Vosburg&Kumar 2004), (Rahm&Hai Do 2000), (Osborne 2013). Kvartaalien liikevaihdon erittely menee tuolloin varmasti sekaisin, ellei näitä asioita huomioida kantojen integraatiovaiheessa. Ongelmia voidaan pyrkiä

minimoimaan ohjelmistoilla, jotka on suunniteltu yhdistämään yrityksen tietovarantoja(Vosburg&Kumar 2004), tai asettamalla (moderneille) tietokannoille skeematason rajoituksia.

Patil & Kulkarni (2012),Rahm&Hai Do (2000) ja Hernández ym. (1998) listaavat tietovaraston siivoamiseen apusovellusten käyttämisen. Edellämainitut ovat samaa mieltä siitä, että suurin työ hoituu vain manuaalisesti kantoja läpikäymällä. Esimerkiksi otejoukkoa voidaan käsitellä suurien datamassojen siivoamiseen tehdyillä sovelluksilla, joilla voidaan löytää poikkeamia ja muita ongelmia otejoukosta. Suurin osa havaituista ongelmista kuitenkin vaatii ihmisen tekemiä korjauksia, esimerkiksi toisteisten esiintymien seulonnassa. Kun otejoukko on siivottu apusovelluksilla, voidaan BI-ratkaisussa se visualisoida. Visualisoinnissa havaitaan usein ongelmia esimerkiksi paikkadataa sisältävissä ja koskevissa esiintymissä.

3.3 Tietokantojen yhdistämisen ongelmia tietovarastossa

Tietokantoja yhdistettäessä, yksikkötasolla korjaamatta jääneet ongelmat moninkertaistuvat. Sovellukset(Vosburg&Kumar 2004) joiden avulla kantoja voidaan yhdistää, eivät välttämättä löydä samaa henkilöä koskevia esiintymiä. Esimerkiksi kirjoitusvirhe (Kuvio 2) yksikkötasolla aiheuttaisi yhdistettyjen kantojen tasolla useamman esiintymän kustakin käsitellyn otejoukon *"Customers"* - attribuutin henkilörivistä. Näitä virheitä jos ajatellaan asiakkaina, joiden kanssa käyty liikevaihto jakautuu useampaan raportin kohtaan, saatetaan luoda virheellistä tietoa esimerkiksi yksiköiden tuottavuudesta.

Toisteisia esiintymiä löydettäessä, voidaan niiden tiedot yhdistää ja kirjata toiminta yhden esiintymän alle. Tuolloin yksittäisen esiintymän sisältämä tieto on täydellisempää kuin esiintymien, josta se on johdettu. Osborne (2013) ja Rahm&Hai Do (2000) toteavat, että täydennetty data tulee myös palauttaa takaisin tietokantoihin, joista yksittäinen esiintymä oli johdettu. Tuolloin näitä kantoja käyttävät ohjelmat voivat hyötyä täydennetystä datasta.

3.4 Tyhjien arvojen käsittely ja tutkimuskysymysten asetus

Käsiteltävässä datassa on tyypillistä, että esiintymiltä löytyy tyhjiä arvoja. Ensinäkemältä puuttuva data tulkitaan huonoksi asiaksi, jota se usein on. Kuitenkin puuttuvaa dataa voidaan tulkita, kuten hiljaisuutta musiikissa, sillä siinä voi piillä paljon potentiaalia. Osborne (2013) mukaan, puutteellisesta datasta voidaan vetää hyviä johtopäätöksiä, jos haettava kysymys muotoillaan oikein.

Yksi tyypillinen tapa (Osborne 2013) mukaan hyödyntää puutteellista dataa on muuttaa nimittäjää, jolla leikataan puutteellisen datan esintymät otejoukosta. Esimerkiksi tutkittaessa otejoukkoa naimisiin menneistä ihmisistä, voitaisiin poissulkea naimattomat henkilöt joukosta, rajaamalla otejoukkoa avioliiton pituuden perusteella. Tuolloin voimme esittää suhteellisen yksinkertaisia kysymyksiä, kuten ”Kuinka moni väestöstä on naimisissa?” Vaihtoehtoisesti voitaisiin myös kysyä keskimääräistä avioliiton pituutta.

Tuolloin puutteellisen datan hyötykäyttö jää kuitenkin pieneksi. Samaista väestödataa käsitellessä, voitaisiin hyvin hyödyntää (avioliitonPituus=”null”) avaimien tarjoamia mahdollisuuksia kysymällä, "Mitä ennusteita on sille, että henkilö on naimisissa?"Tässä voitaisiin tarkastella esimerkiksi iän ja ammatin vaikutuksia asetettuun tutkimuskysymykseen. Vaihtoehtoisesti samaa ideologiaa voidaan hyödyntää esimerkiksi tutkittaessa eronneiden avioliittojen keskimääräistä kestoa ja siihen vaikuttavia tekijöitä, kuten ikä, sijainti, tai ammatti. Tällöin puuttuvaa dataa voidaan hyödyntää esimerkiksi liiketoiminnassa, tutkittaessa päättyneitä liiketoimintasuhteita asiakkaiden kanssa. Jo olemassa olevia liiketoimintasuhteita ei tuolloin ole mielekästä sisällyttää tarkasteluun (Osborne 2013).

Puuttuvan datan hyötykäyttö toimii vain otejoukoissa, joissa puutteellinen data on tarkoituksellista ja käsiteltävä data pitää paikkansa. Järkevissä BI-ratkaisuissa otejoukkoja voidaan leikata edellämainitun esimerkin tapaisesti, vaikkapa tarkastellessa yhtiön tulosta tietyllä paikkakunnalla. Tampereella toimivan osaston tuottojen analyysissä on epämielekästä sisällyttää PK-seudun osastojen liiketoimintaa analyysiin. Yksi tapa toteuttaa tätä, on esimerkiksi Microsoftin PowerBI:ssä rajoittimia

käyttämällä.

Rajoittimien avulla voidaan tutkia tarkemmin havaittuja yhtäläisyyksiä, tai poikkeamia saman raportointinäkymän sisällä. Esimerkiksi sijainnin, managerin, tai vaikka markkinoinnin tehokkuutta voidaan tutkia eri ajanjaksoina rajoittimien avulla. Lisäämällä useampia rajoittimia, voidaan tutkittavaa kohdejoukkoa rajata esimerkiksi katkenneisiin liiketoimintasuhteisiin, osastoihin, tai huoltotehtävien määrään.

3.5 Virheelliset ja puuttuvat arvot

Tietokannoista ja varastoista löytyy kuitenkin usein esiintymiä, joissa dataa puuttuu tai se on epätäydellistä. Suuria tietovarastoja käsitellessä, on jokaisen esiintymän ja niissä esiintyvien virheiden läpikäynti manuaalisesti äärimmäisen epätehokasta. Virheiden löytämiseen voidaan käyttää apuna siihen tarkoitettuja sovelluksia (Patil&Kulkarni 2012) ETL-prosessin aikana.

Puuttuvien arvojen käsittelyssä helpoin tapa on ollut jättää arvo huomioimatta, jos siinä on puutteita (Schafer & Olsen 1998). Monimuuttuja-analyysissä (engl. *Multivariable analysis*) havaitaan usein puuttuvia arvoja. Schafer&Olsen (1998) käyttää esimerkkinä pitkittäistutkimusta, jossa tutkimukseen osallistuvia tahoja saattaa pudota pois. Näiden tahojen tuottama data jää huomioimatta, jos tyhjiä arvoja sisältävät rivit hylätään otejoukosta. Vaihtoehtoisesti tätäkin dataa voidaan hyödyntää, etsimällä yhtäläisyyksiä tyhjiä arvoja sisältävien tahojen välillä kohdan 3.4 esittämällä tavalla.

Moni-imputointi -testauksessa (Schafer&Olsen 1998) (engl.*Multiple-Imputation*) käsiteltävälle otejoukolle pyritään löytämään todennäköisiä arvoja, puuttuvien arvojen tilalle. Tyhjiä arvoja sisältäville soluille matriisissa voidaan löytää todennäköiset arvot ennustejakauman avulla *(ts. prediktiivinen jakauma)*. BI-ratkaisujen usein tukiessa Pythonia-, tai R-kieltä, voidaan raportin sisälle kirjoittaa algoritmi ennustejakauman laskemiselle. Ennustejakauman perusteella voidaan syöttää arvioituja lukuja haluttuihin soluihin (Schafer&Olsen 1998). Moni-imputointi -testauksessa kuitenkin samasta otejoukosta luodaan useampi jatkettu otejoukko. Näiden luomisessa

voidaan käyttää esimerkiksi leikattua otejoukkoa eri keskihajontaluvuilla, tai edellisiä otejoukkoja. Luotujen otejoukkojen pohjalta tehdään lineaarinen regressioanalyysi, jonka pohjalta luodaan menestynein implementaatio, eli todennäköisin ennuste.

Tarkempaan analytiikkaan pyrkiessä, voidaan täydentää moni-imputointi -testauksen (Schafer&Olsen 1998) ideaa, Srisakaokulin (2017) ehdottamalla moni-implementaatio -testauksella. Tuolloin Schaferin 1998 ehdottamaa mallia täydennetään muilla tekniikoilla, kuten kNN (k-Lähin-Naapuri) ja NB (Naive-Bayes):in menetelmillä. Testauksessa luoduilla otejoukoilla opetetaan koneoppimisalgoritmia, jonka tuottamaa analyysiä voidaan käyttää BI-raportissa ennusteena. Prosessi kuitenkin vaatii paljon laskentatehoa ja tämä tulisi ottaa huomioon otejoukon läpivientikertoja pohtiessa. Esimerkiksi kNN algoritmi on hyvin raskas, ulottuvuuksien ja otejoukon koon kasvaessa. Algoritmi toimii siten, että esimerkiksi huomisen ennustettua huoltotöiden määrää etsiessä, sille annetaan tähän mennessä tuotettu huoltotyöhistoria. Algoritmi luo tästä otejoukosta matriisin ulotteisuudessa ”x”. Ennuste tästä otejoukosta saadaan syöttämällä uusi esiintymä otejoukkoon, esimerkiksi viikonpäivän, päivämäärän ja sään perusteella. Algoritmi laskee tämän jälkeen ”k” määrän lähimpiä esiintymiä matriisissa. Etäisyys näihin esiintymiin lasketaan euklidisen avaruuden vektorien pituudesta. Laskettaessa esimerkiksi kymmenen lähintä esiintymää, uuden esiintymän arvo huoltotöiden ennustetusta lukumäärästä voidaan johtaa näiden keskiarvosta, tai valitsemalla useimmin toistuva luku. Tarkasteltaviksi ennakkotiedoiksi voidaan ottaa esimerkiksi lähin viikonpäivä muutamana aiempana vuotena ja viikkoina, menneiden lähipäivien säätila ja ennustettu säätila seuraavalle päivälle. K-arvon ja otejoukon kasvaessa, laskenta hidastuu.

Yhdistämällä tähän muita algoritmeja, tai eri parametreilla luotavia kNN-ennusteita, luodaan yksinkertainen eteenpäin kytketty neuroverkko. Tässä kolmi-/nelitasoisessa neuroverkossa halutut parametrit voidaan löytää syötettyjen matriisien attribuuteista. Näistä esimerkiksi lähtöneuroneina voidaan pitää mainittuja tietokantoja, jotka sisältävät huoltotöiden historian ja säädatan. Nämä muodostavat neuroverkon ensimmäisen kerroksen (Srisakaokul,Wu, Astorga 2017). Välikerroksien neuroneja

ovat ajettavat algoritmit ja loppuneroni on etsitty lukuarvo tai riippuen toteutuksesta, lukuarvot ja etsityt päivät. Näin pyritään löytämään mahdollisimman täydellinen tulosjoukko BI-raportille visualisoitavaksi.

4 Ratkaisumallit

Kappaleessa käydään läpi ETL-prosessin askeleet, ennusteet, BI-ratkaisun ulkoasuun littyviä seikkoja ja mainitaan alalla tunnettuja palveluntarjoajia BI-ratkaisujen tutottamiseen.

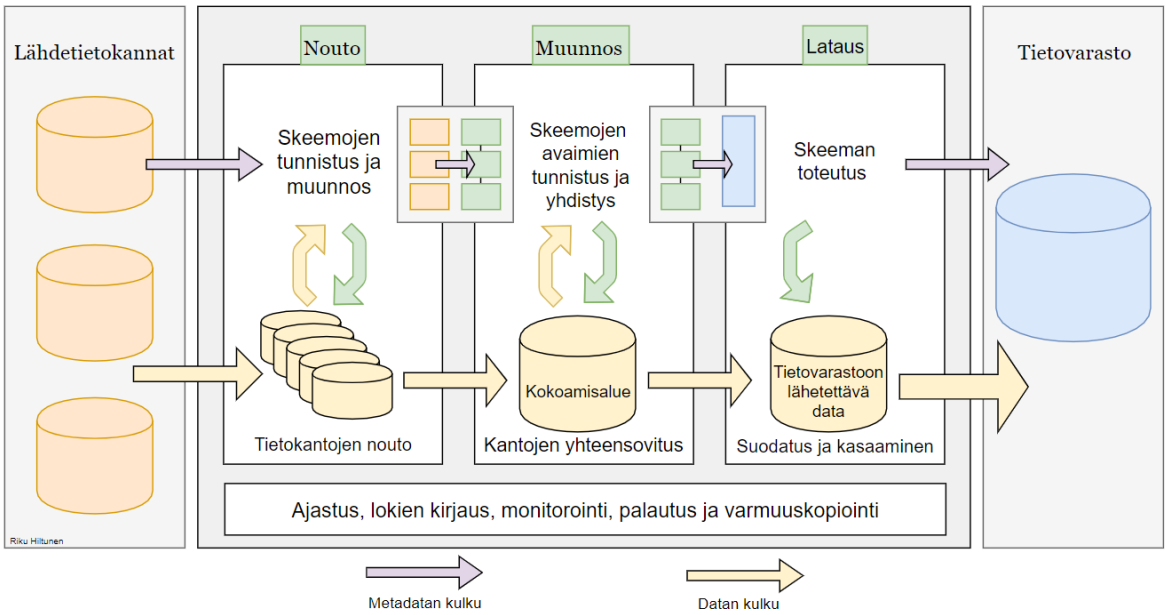
4.1 ETL-prosessin askeleet likaisen datan poistamisessa

Nouto-, muunnos- ja lataus-vaiheista koostuvaa prosessia kutsutaan ETL-prosessiksi ("engl. *Extracting-Transformation-Loading*). Kuviosta 3 nähdään, miten ETL-prosessin mukaisesti dataa ja likaisen datan syntymistä voidaan estää tietovarastossa (Rahm&Hai Do 2000).

4.1.1 Nouto

Rahm ja Hai Do (2000) nimeävät nouto (engl.*"Extraction"*) -vaiheen ensimmäiseksi vaiheeksi ETL-prosessissa. Tässä vaiheessa noudetaan otejoukkoja lähdetietokannoista. Tietokannat voivat olla yrityksen, asiakkaiden, tai kolmannen osapuolen omistamia. Datan käytettävyyttä voidaan parantaa tässä vaiheessa tehokkaasti sovellusten avulla (Patil&Kulkarni 2012). Sovellukset osoittautuvat hyödyllisimmiksi puuttuvaa dataa, toisteisuutta ja syötevirheellistä dataa havaitessa ja siivotessa (Patil&Kulkarni 2012). Toisinsanoen yksikkötason virheiden korjauksessa. Usein iso osa työstä jää kuitenkin matalan tason ohjelmille, tai manuaalisesti hoidettaviksi (Rahm&Hai Do 2000).

Patil ja Kulkarni (2012) nimittävät vaihetta myös data-analyysin vaiheeksi. Tavoitteena tässä vaiheessa on saada metatietoa käsiteltävistä otejoukoista, hahmottaa niiden välisiä suhteita ja käytettävyyttä. Dataa usein käsitellään ositettuna erillisellä tallennusauleella, prosessin nopeuttamiseksi (Rahm&Hai Do 2000).



Kuvio 3. ETL - kaavio tietovarastosta. Rahm&Hai Do (2000)

4.1.2 Muunnos

Rahm ja Hai Do (2000) nimeävät ”Transformation” (suom.*”Muunnos”*) -vaiheen seuraavaksi vaiheeksi ETL-prosessissa. Tässä vaiheessa sovitellaan tietokantojen skeemoja ja integroidaan nämä yhteen. Tavoitteena on päästä tilanteeseen, jossa edellisessä vaiheessa siivottujen otejoukkojen data saadaan sovitettua yhdeksi otejoukoksi. Tuolloin yhden otejoukon esiintymät ovat täydellisempiä, jos osajoukkojen sisältämä data saadaan yhdistettyä otejoukon esiintymissä.

Rahm ja Hai Do (2000) listaavat tämän vaiheen tehtäviin kuuluvan myös sääntöjen ja työjärjestyksen tekemisen tietokantojen välille. Käsitellessä heterogeenisiä tietovarastoja, samaa tietoa saattaa olla toisteisena eri kannoissa, jolloin näiden välille tulee rakentaa riippuvuussuhteet ja päättää, mitä avain-arvo-pareja käytetään massadatan identifioinnissa (Patil&Kulkarni 2012). Omistetusta datasta saadaan tällöin parempilaatuista ja sen kokoa saadaan pienennettyä toisteisuuden vähentyessä (Rahm&Hai Do 2000).

4.1.3 Lataus

Kolmantena vaiheena Rahm ja Hai Do (2000), sekä Patil ja Kulkarni (2012) listaavat tietojen varmennuksen. BI-toteutuksissa tätä vaihetta voidaan pitää vaiheena, jossa käsiteltävää otejoukkoa visualisoidaan, tai sitä viedään muihin sovelluksiin analysoitavaksi. Rahm ja Hai Do (2000) toteavat, että otejoukon ollessa käytettävissä, tulee siitä vedettäviä johtopäätöksiä analysoida ja varmistaa niiden oikeellisuus. Tässä vaiheessa keskitytään siis virheiden löytämiseen ja korjaamiseen luodusta otejoukosta.

4.1.4 Tietovarastointi

Rahm ja Hai Do (2000) nimeävät viimeiseksi vaiheeksi tuotetun otejoukon viemisen lähdetietokantoihin ja yrityksen tietovarastoon. Tässä vaiheessa tavoitteena on varmistaa, että ETL-prosessissa toteuteut toimenpiteet ja luodut skeemat toimivat oikein. BI-ratkaisua tuottaessa, oheishyötynä voidaan nähdä yrityksen omistaman massadatan laadun nousu. Tätä dataa voidaan mahdollisesti hyödyntää myös muissa sovelluksissa, jos nämä lukevat luotua skeemaa (Rahm&Hai Do 2000).

Patil ja Kulkarni (2012) kirjoittavat, että lähetetty data tulee noutaa uudelleen käsittelyyn. Tuolloin nähdään, jos ETL-prosessin automatisointi on mennyt oikein, kun käsiteltävä data päivitetään, eli uudelleenkäsitellään (Rahm&Hai Do 2000). Tavoitteena on ihannetila, jossa tietovarasto kykenee päivittämään itsensä luotettavasti, minimaalisella määrällä virheitä. Lähdetietokantojen päivittyessä jatkuvasti, on niistä löytyvien virheiden määrä vain ajan ja ETL-prosessin uudistamisen kysymys.

4.2 Ulkoasu ja visuaalinen suunnittelu

BI-toteutuksessa suurena tekijänä toimii myös sen ulkoasun suunnittelu. Tiivistäen Few (2006) teoriaa, suunnittelun tulisi alkaa raportointinäkymän tavoitteista. Sen tulisi kertoa monimutkainen kokonaisuus yksinkertaisesti. Tässä suuressa osassa toimii sen visuaalinen asettelu, raportointinäkymän selkeyden riippuessa tästä. Tyypillinen virhe toteutuksissa on liian monen elementin käyttö samalla sivulla. Tämä

tekee raportointinäkymästä hankalan ymmärtää (Few 2006). Nyrkkisääntönä voisi pitää maksimissaan 5-9 elementtiä sivulla. Porautuminen on hyvä työkalu, jos monimuotoisuutta halutaan tarjota raportointinäkymässä, käytetyn BI-alustan tätä tukiessa.

Raportointinäkymässä käyttäjän tulisi saada selville yleisimmät raportin tarjoamat vastaukset noin viidessä sekunnissa. Tässä hyvä työkalu on tekstielementti, joka näyttää muutoksen vertailuajankohtaan vaikkapa prosentteina. Nämä asemoidaan raportointinäkymässä omaksi ryppääkseen, käyttäen suunnittelussa vaikkapa käänteistä pyramidia (Few 2006). Tämän mukaan tieto kategorisoidaan kolmeen osa-alueeseen, huomionarvoisiin asioihin, jotka esimerkiksi kuvaavat lyhyen aikavälin muutosta; trendeihin, joilla visualisoidaan edellämainitut elementit ja taustatietoon, jossa kerrotaan, mihin edellämainitut perustuvat. Käänteisen pyramidin teoria on tuttu journalistiikasta. Journalistiikan tavoitteena on kertoa tarinaa, kuten BI-ratkaisuilla on tavoitteena kertoa yrityksen tarina.

4.3 Ennusteet

Koneoppimisen avulla voidaan BI-palvelulle antaa lisäarvoa. Jotkin BI-palvelut tarjoavat ennusteita sisäänrakennettuna palveluunsa, esimerkiksi Tableaussa. Useimmissa, kuten Microsoftin PowerBI:ssä, on mahdollista syöttää omia koneoppimisalgoritmeja, joilla BI:llä tuotetun raportin laatua voidaan parantaa. Edellämainitussa voidaan esimerkiksi kirjoittaa Pythonilla haluttujen koneoppimisalgoritmien tuottamia tuloksia raporttiin (Cofsky 2018).

Srisakaokul,Wu, Astorga (2017) nostavat ensimmäiseksi ongelmaksi koneoppimisalgoritmeissa epätäydellisen datan. ETL-prosessissa(Rahm&Hai Do 2000) siivotut otejoukot ovat tuolloin hyvä pohja koneoppimiselle. Varsinkin jos ennusteille nähdään lisäarvoa ja data on jo käsitelty kohdassa 3.5 esitetyillä tavoilla. (Srisakaokul,Wu, Astorga 2017) Mukaan koneoppimisalgoritmien toisena yleisenä kompastuskivenä on se, miten siitä saatavat tulokset saadaan esitettyä järkevästi (Locklin 2014). Tähänkin tarpeeseen BI vastaa hyvin. Koneoppimisalgoritmin menestystä

mittaa kuitenkin pitkälle se, kuinka hyvin sen tuottamat tulokset saadaan kommunikoitua lukijalle tehokkaasti.

5 Yhteenveto

Tutkielmassani esitän ongelmien pääpiirteet, jotka tulee ottaa huomioon BI-ratkaisuja tehdessä. Ensimmäisessä kappaleessa läpikäyn, mitä BI-ratkaisuilla tarkoitetaan. Nojaan omien kokemuksieni pohjalle työelämästä ja esitän alan kirjallisuuden toteamat normit esimerkein. Keskityin tutkielmassa myös tietokantojen yhdistämisessä havaittuihin ongelmiin, yksikkö-, ja skeematasolla. Näistä voidaan esimerkkeinä antaa toisteiset esiintymät ja vapaan syötteen kenttien esiintymät tietokannoissa. Yhdistin tämän jälkeen näissä koetut ongelmat tietovarastojen yhdistämisen ongelmiin.

Käytännön BI - ratkaisuissa joudutaan usein yhdistämään useamman tietokannan dataa. Näistä yhtenä esimerkkinä mainittakoon vaikka paikkadata. Sitä esiintyy esimerkiksi laiterekistereissä ja kiinteistöissä. Usein näistä saatu data halutaan visualisoida kartalle. BI-alustat eivät osaa piirtää dataa, jonka paikkatietona on "Joensuun toimisto". Tämä saattaa aiheuttaa yllättävän paljon työtä, sillä vastaavien esiintymien selvittelyssä tulee löytää kyseisen toimiston osoite. Näiden ongelmien ratkaisuun kerron ETL-prosessista ja sen mukaisesta tietokantojen ja tietovarastojen käsittelystä.

Tutkielmassa käsitellään myös puuttuvien arvojen täydentämistä otejoukkoihin ja niistä luotavia ennusteita. Näistä mainitaan muutama, kuten kNN ja Naive-Bayes. Algoritmeista kNN:ää käsitellään enemmän ja käydään lyhyt esimerkki kNN:n pohjalta, kuinka ennusteita voidaan luoda BI-ratkaisujen käytettäväksi. Jatkotutkimusta voitaisiin toteuttaa BI-ratkaisuja tarjoavien palveluiden eroista. Näistä voitaisiin mainita esimerkiksi PowerBI, Tableau, Qlik, SiSense ja IBM Cognos. Muita jatkotutkimuksen aiheita voisivat olla hyvät käytänteet visualisoinneissa ja tietovarastojen optimointimetodit.

Kirjallisuutta

Chen H, Roger H. L. Chiang, Veda C. Storey 2012. *MIS QUARTERLY: Business Intelligence and Analytics: From Big Data to Big Impact*. Lehdessä:MIS Quarterly (Joulukuu/2012) Voluumi:36 Numero:4 Sivunumero:1165-1188 DOI: 10.2307/41703503 Saatavilla WWW-muodossa <URL: https://pdfs.semanticscholar.org/f5fe/b79e04b2e7b61d17a6df79a44faf358e60cd.pdf>. Viitattu 06.11.2018.

Amanda Cofsky 2018. *Power BI Desktop August 2018 Feature Summary*. Saatavilla WWW-muodossa <URL: https://powerbi.microsoft.com/en-us/blog/power-bi-desktop-august-2018-feature-summary/>. Viitattu 08.12.2018

Columbus, L. 2017. *53(precent) Of Companies Are Adopting Big Data Analytics*. Julkaistu: Forbes, 24.12.2017 Saatavilla WWW-muodossa<URL: https://www.forbes.com/sites/louiscolumbus/2017/12/24/53-of-companies-are-adopting-big-data-analytics/#d133b5539a19>. Viitattu 4.11.2018.

- tekninen raportti, Dresner Advisory Services Dresner Advisory Services. 2017. *Dresner Advisory Services Publishes 2017 Big Data Analytics Market Study*. Julkaisija: Dresner Advisory Services. Saatavilla WWW-muodossa <URL: http://www.bigdataanalyticsreport.com/>. Viitattu 6.11.2018.

Stephen Few 2006. *Information Dashboard Design: The Effective Visual Communication of Data*. ISBN:0596100167 Julkaisija: O'Reilly Media, Inc. ©2006 Saatavilla WWW-muodossa <URL: https://the-eye.eu/public/Books/IT%20Various/information_dashboard_design.pdf>. Viitattu 08.12.2018

Bhumika Hansoti 2010. *Business Intelligence Dashboard in Decision Making*. Julkaisija: College of Technology Directed Projects, 15 Saatavilla WWW-muodossa<URL: https://docs.lib.purdue.edu/techdirproj/15/>. Viitattu 08.12.2018

Hernández, M. 1998. *Data Mining and Knowledge discovery: Volume 2: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem*. Julkaisija: Data Mining and Knowledge Discovery DOI:10.1023/A:1009761603038. Saatavilla

WWW-muodossa <URL: https://www.semanticscholar.org/paper/Real-world-Data-is-Dirty%3A-Data-Cleansing-and-The-Hern%C3%A1ndez-Stolfo/000bd5997cc64c6c2f48d93202e7998cbc49479a>. Viitattu 11.11.2018.

Pro Gradu - tutkielma, Jyväskylän Yliopisto. Hokkanen, M. 2012. *Business Intelligence-muutoksen paradigma johdon laskentatoimessa*. Julkaisija: Jyväskylän Yliopisto. Saatavilla WWW-muodossa <URL: http://urn.fi/URN:NBN:fi:jyu-201211213053>. Viitattu 6.11.2018

- Tekninen raportti, IBM IBM 2011. *IBM Tech Trends report*. Julkaisija: IBM. Saatavilla WWW-muodossa<URL: https://ai.arizona.edu/sites/ai/files/MIS510/2011ibmtechtrendsreport.pdf>. Viitattu 6.11.2018.

Pro Gradu - tutkielma, Jyväskylän Yliopisto. Korhonen, J. 2017. *Korhonen, J. ulkoinen liiketoimintatiedon hallinta mielikuvien ja maineen taloudessa* . Julkaisija: Jyväskylän Yliopisto. Saatavilla WWW-muodossa <URL: http://urn.fi/URN:NBN:fi:jyu-201206191900>. Viitattu 6.11.2018.

Scott Locklin 2014. *Neglected machine learning ideas*. Saatavilla WWW-muodossa <URL: https://scottlocklin.wordpress.com/2014/07/22/neglected-machine-learning-ideas/>. Viitattu 08.12.2018

Osborne,J. 2013. *Best practices in data cleaning : a complete guide to everything you need to do before and after collecting your data*. Kustantaja: Sage. Saatavilla WWW-muodossa <URL:https://www.researchgate.net/publication/266714997_Best_practices_in_data_cleaning_A_Complete_Guide_to_Everything_You_Need_to_Do_Before_and_After_Collecting_Your_Data>. Viitattu 1.1.2004.

Patil, R. Kulkami, R.V. 2012. *A Review of Data Cleaning Algorithms for Data Warehouse Systems*. Julkaistu: International Journal of Computer Science and Information Technologies, Vol. 3 (5) , 2012, sivut:5212 - 5214 Saatavilla WWW-muodossa <URL: http://ijcsit.com/docs/Volume/203#vol3issue5/ijcsit2012030556.pdf>. Viitattu 17.11.2018.

.

Rahm, Erhard, and Hong Hai Do 2000. *Data cleaning: Problems and current*

approaches. Julkaistu: IEEE Data Engineering Bulletin 23(4): 3–13. Saatavilla WWW-muodossa <URL: https://www.betterevaluation.org/sites/default/files/data_cleaning.pdf>. Viitattu 11.11.2018.

Sahay, A. 2017. *Data Visualization, Volume II : Uncovering the Hidden Pattern in Data Using Basic and New Quality Tools*. Julkaisija: Business Expert Press. ISBN: 9781631577321. Saatavilla WWW-muodossa <URL: https://www.businessexpertpress.com/books/data-visualization-volume-ii-uncovering-hidden-pattern-data-using-b Viitattu 6.11.2018.

Joseph Schafer, Maren Olsen 1998. *Multiple imputation for multivariate missing-data problems: a data analyst's perspective*. Julkaisu:Multivariate behavioral research, 33 4, 545-71 . PMID: 26753828 DOI: 10.1207/s15327906mbr33045 Saatavilla WWW-muodossa <URL: https://spdfs.semanticscholar.org/02f6/ad13ff5f2976047341eb497b14f72754c747.pdf >. Viitattu 10.12.2018

Sherman, R. 2012. *SHERMAN, R., 2014. Business Intelligence Guidebook : From Data Integration to Analytics. San Francisco: Elsevier Science and Technology.*. Julkaisija: Elsevier Science & Technology. ISBN: 9780124115286. Saatavilla WWW-muodossa <URL: https://www.elsevier.com/books/business-intelligence-guidebook/sherman/978-0-12-411461-6>. Viitattu 4.11.2018.

Srisakaokul, ym. *Multiple-Implementation Testing of Supervised Learning Software*. URI: http://hdl.handle.net/2142/91645 Julkaistu: IDEALS(Illinois Digital Environment for Access to Learning and Scholarship) 2016-10-10. AAAI-18 Workshop on Engineering Dependable and Secure Machine Learning Systems Saatavilla WWW-muodossa <URL: http://taoxie.cs.illinois.edu/publications/edsmls18-mitest.pdf>. Viitattu 08.12.2018.

Vosburg, J. Kumar, A 2001. *Managing Dirty data in organizations using ERP: lessons from a case study*. Lehdessä: Industrial Management & Data Systems, Vol. 101 Issue: 1, pp.21-31 Julkaisija: MCB UP Ltd DOI: <https://doi.org/10.1108/02635570110365970> Saatavilla WWW-muodossa <URL: https://www.emeraldinsight.com/doi/abs/10.1108/

02635570110365970>. Viitattu 11.11.2018.