Open your mind. LUT.
Lappeenranta University of Technology

TUOTANTOTALOUDEN KOULUTUSOHJELMA

Data-analytiikka rahanpesun estämisessä

Data-analytics in anti-money laundering

Kandidaatintyö

Aaro Hassinen

TIIVISTELMÄ

Tekijä: Aaro Hassinen

Työn nimi: Data-analytiikka rahanpesun estämisessä

Vuosi:2019 **Paikka: Lappeenranta**

Kandidaatintyö. Lappeenrannan teknillinen yliopisto, tuotantotalous.
47 sivua, 5 kuvaa ja 3 taulukkoa ja 1 liite
Tarkastaja(t): Tutkijatohtori, TkT Lasse Metso

Hakusanat: Data-analytiikka, Rahanpesu, Rahanpesun estäminen
Keywords: Data-analytics, Anti-money Laundering, AML

Data-analytiikan käyttö yleistyy yrityksissä ja sen avulla pyritään saamaan kilpailuetua. Työn tavoitteena on selvittää, miten finanssialan organisaatio pystyy hyödyntämään data-analytiikkaa ja erityisesti, miten finanssialan organisaatio pystyy käyttämään data-analytiikkaa rahanpesun estämisessä ja millaisia haasteita se aiheuttaa. Työn aihepiireihin tutustutaan tieteellisten kirjojen ja artikkelin avulla, sekä neljään eri finanssialan organisaatioon toteutetun haastattelututkimuksen avulla.

Data-analytiikan käyttökohteita on useita ja se soveltuu erityisesti korvaamaan manuaalista työtä, tehostamaan prosesseja ja vähentämään kustannuksia. Haasteina data-analytiikan hyödyntämisestä finanssialan organisaatiossa tulee saatavilla olevasta datasta ja sen laadusta. Työssä havaittiin, että data-analytiikkaa hyödynnetään jo monessa finanssialan organisaatiossa ja ei pelkästään rahanpesun estämisessä. Data-analytiikkaa hyödynnetään rahanpesun estämisen lisäksi sisäisissä prosesseissa ja rahanpesun estämisen tukena käytetään data-analytiikka ohjelmistoja. Työssä löydettiin yhteneviä ongelmia data-analytiikan hyödyntämisessä kirjallisuudesta ja haastateltavista finanssialan organisaatioista.

SISÄLLYSLUETTELO

1 JOHDANTO

Rahanpesu on merkittävä ongelma talouksille ja finanssialan toimijoille ympäri maailmaa. Globalisaation, informaation ja teknologian kehityksen takia rahan liikkuminen maasta toiseen on nopeampaa ja helpompaa, kuin koskaan ennen ja tämä tekee rahanpesun estämisestä entistä vaikeampaa.

Rahanpesun estäminen on erittäin tehokaskeino estää järjestäytynyttä rikollisuutta. Järjestäytyneestä rikollisesta toiminnasta saatu taloudellinen hyöty on haitaksi terveelle talouden kehitykselle ja samalla rahoituslaitosten luotettavuus ja uskottavuus on vaarassa, jos rikolliset pystyvät peittämään rikollisesta toiminnasta saadun taloudellisen hyödyn ja käyttämään näitä rahoja esimerkiksi terrorismin tukemiseen. (HE 25/2008)

UNODC:n (United Nations Office on Drugs and Crime) tekemän tutkimuksen mukaan vuonna 2009 rikolliset ja erityisesti huumeiden salakuljettavat ovat saattanet pestä rahaa jopa 1600 miljardin dollarin edestä. Tämä summa vastaa 2.7% koko maailman bruttokansantuotteesta. Tutkimuksessa todetaan, että kun laiton raha on päässyt osaksi globaalia talousmarkkinaa, sen alkuperää on entistä vaikeampi selvittää ja se vain kannustaa järjestäytynyttä rikollisuutta jatkamaan toimintaa. (UNODC, 2011)

Suomi on toiminut FATF:n jäsenenä vuodesta 1991 lähtien. FATF (Financial Action Task Force) tekee työtä rahanpesua ja terrorismin rahoittamista vastaan. FATF kehittää ja antaa jäsenmaille suosituksia, kuinka toimia rahanpesua ja terrorismin rahoittamista vastaan. FATF:n jäsenmaat ovat sitoutuneet poliittisesti noudattamaan heidän suosituksiansa ja niitä valvotaan vuosittain kyselyillä ja arvioinneilla. FATF on julkaissut 1990 vuodesta alkaen 40 suositusta rahanpesun ja terrorismin rahoituksen vastaiseen toimintaan. FATF on suorittanut Suomeen maa-arvioinnin viimeksi vuonna 2019 ja sen tulokset on julkaistu 16.4.2019. (Valtiovarainministeriö, 2019)

Data-analytiikkaa on tieteellinen prosessi, jossa saadaan saatavilla olevasta datasta hyödyllistä informaatiota päätöksenteon tueksi. Nykypäivänä datan määrä on kasvanut erittäin suureksi. Nykytahdilla maailmassa syntyy 2.5 kvintiilitavun edestä, eli 2.5 miljoonan teratavun edestä dataa (Marr 2018). Dataa syntyy niin paljon, että sitä on vaikea edes suhteuttaa mihinkään. 2.5 miljoonaa teratavua on verrattavissa 500 miljoonaan normaalilla älypuhelimella otettuun

valokuvaan. Data-analytiikkaa hyödynnetään kaivautumaan tähän tiedon määrään ja etsitään sieltä kuvioita, trendejä ja tehdään ennusteita saatavilla olevan tiedon perusteella. (Gupta 2016, s. 1)

1.1 Työn tavoitteet ja tutkimuskysymykset

Tämän työn tavoitteena on saada luotua kattava selvitys nykytilanteesta, kuinka data-analytiikkaa hyödynnetään rahanpesun estämisessä tällä hetkellä. Tämän lisäksi tarkoituksena on selvittää tulevaisuuden mahdollisuudet data-analytiikalle tässä aihealueessa. Työ tehdään toimeksiantona finanssialan organisaatiolle ja tärkeänä on päästä sellaiseen johtopäätökseen, jossa selviäisi mitä data-analytiikan keinoja olisi mahdollista käyttää rahanpesun estämisessä ja mitä lisäarvoa siitä voitaisiin saada. Tämän kandidaatintyön tutkimuskysymykset ovat seuraavat:

1. Millaista lisäarvoa data-analytiikan hyödyntäminen voi tuoda finanssialan organisaatiolle?
2. Miten data-analytiikkaa voidaan hyödyntää rahanpesun estämisessä?

1.2 Tutkimusmenetelmät ja rajaus

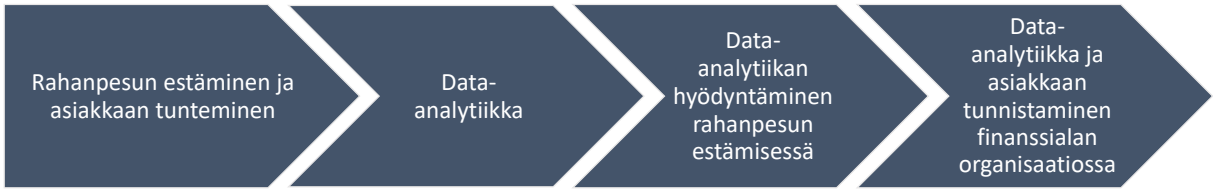
Työ tehdään kirjallisuuskatsauksena. Työtä varten haastateltiin alan ammattilaisia tutkimusta varten ja kysyttiin heidän tavoitteistaan ja kuinka he kokevat data-analytiikan mahdollisuudet ja mitä heidän mielestänsä data-analytiikan avulla voi saavuttaa. Haastatteluista saatua tietoa hyödynnettiin täydentävänä tietona kirjallisuuskatsauksen ohelle. Samalla selvitettiin, millaista dataa asiakkaista on käytettävissä ja millaisia haasteita datan suhteen heillä on. Soveltavana osuutena haastattelin neljän eri finanssialan organisaation edustajaa, jotka työskentelevät data-analytiikan ja rahanpesun estämisen parissa. Tämän soveltavan osuuden aineisto on kerätty puolistrukturoidun haastattelututkimuksen avulla. Puolistrukturoidussa haastattelussa kaikille haastateltaville esitetään samat kysymykset ja kysymyksien vastauksia ei sidota valmiisiin

vastausvaihtoehtoihin (Hirsjärvi & Hurme 2015, s.47). Haastattelun kysymykset (Liite 1) lähetettiin haastateltaville sähköpostitse.

Työn aihepiirinä on tutkia data-analytiikan menetelmiä ja tekniikoita, joita voitaisiin hyödyntää rahanpesun estämisessä ja rahanpesun estämisen tuomissa haasteissa. Työ rajattiin koskemaan data-analytiikan keskeisiä aihealueita, eikä syventyä vain yhteen aihealueeseen. Työn alussa syvennytään data-analytiikkaan ja dataan, koska ne ovat oleellisessa osassa tätä työtä. Tämän jälkeen siirrytään tämän hetkisiin data-analytiikan käyttökohteisiin rahanpesun estämisessä. Viimeisenä luodaan yhtenäinen kokonaisuus, jossa yhdistyy data-analytiikan mahdollisuudet ja suositukset aiheeseen liittyen.

1.3 Työn rakenne

Kuvassa 1 havainnollistetaan tämän työn kulkua pääpiirteittään. Työssä perehdytään ensin rahanpesuun ja sen estämiseen ja asiakkaan tunnistamisprosessiin ja asiakkaan tuntemiseen. Asiaa käsitellään ensisijaisesti lainsäädännölliseltä kannalta. Seuraavaksi työssä perehdytään kirjallisuudesta löytyvään tietoon data-analytiikasta ja sen tuomista hyödyistä ja haasteista. Data-analytiikan käsittelyssä syvennytään dataan, big dataan ja koneoppimiseen. Tämän jälkeen työssä perehdytään, kuinka kirjallisuudessa on tuotu esille data-analytiikan käyttökohteita rahanpesun estämisessä. Työn viimeisessä osuudessa ennen johtopäätöksiä syvennytään työn soveltavaan osuuteen, eli haastatteluihin. Tässä vaiheessa käydään läpi haastateltujen henkilöiden vastaukset ja tuodaan esille yhtenevät ja eriävät vastaukset yhteenvedossa. Johtopäätöksissä käydään läpi vastaukset tutkimuskysymyksiin ja luodaan yhtenäinen kokonaisuus, jossa yhdistyy kirjallisuudesta ja haastatteluista saatu tieto.



Kuva 1. Työn kulku

2 RAHANPESUN ESTÄMINEN JA ASIAKKAAN TUNTEMINEN

Tässä kappaleessa käsitellään mitä rahanpesu tarkoittaa ja kuinka se on määritelty Suomen laissa. Samalla tässä kappaleessa käsitellään, kuinka finanssialantoimijat hoitavat heidän asiakkaan tunnistamisprosessin.

Rahanpesulla tarkoitetaan rikoslain 32 luvun 6-10 §:n mukaista toimintaa. Rahanpesuun syyllistyy taho, joka käyttää, ottaa vastaan, muuntaa, luovuttaa, siirtää, välittää tai pitää hallussaan rikollisella toiminnalla saatua omaisuutta tai pyrkii peittämään rikollisella toiminnalla saadun omaisuuden tai hyödyn laittoman alkuperän tai avustaakseen rikoksen tekijää peittämään omaisuuden tai hyödyn alkuperää. (Rikoslaki 32 luku 6-10 §)
Rahanpesulain tavoitteena on ”estää rahanpesua ja terrorismin rahoittamista, edistää tällaisen toiminnan paljastamista ja selvittämistä sekä tehostaa rikoksen tuottaman hyödyn jäljittämistä ja takaisinsaantia.”(Rahanpesulaki 1 luku 1§)

Suomen valtionvarainministeriö vastaa rahanpesun rahoittamisen riskiarvion luomisesta ja sen toimittamisesta Euroopan komissioon. Riskiarviossa on tunnistettava ja otettava huomioon Suomen rahanpesun ja terrorismin rahoittamisen riskejä ja valtionvarainministeriön on päivitettävä sitä säännöllisesti. Valtionvarainministeriön tuottaman riskiarvion tarkoituksena on”

- Yksilöidä rahanpesun tai terrorismin rahoittamisen riskit toimialoilla;
- Tukea ja tehostaa rahanpesun ja terrorismin rahoittamisen torjuntaa ja voimavarojen kohdentamista;
- Tukea eri toimialojen rahanpesun ja terrorismin rahoittamisen torjuntaa koskevien yhdenmukaisten toimintatapojen laadintaa…
- Kuvata rahanpesun ja terrorismin rahoittamisen torjunnan rakenteita ja yleisiä toimenpiteitä, henkilötyövuosia sekä valtion ja muun julkisen talouden rahoitusta.” (Rahanpesulaki 2 luku 1§)

Ilmoitusvelvollisen on velvollisuus laatia riskiarvio rahanpesun ja terrorismin rahoittamisen riskien tunnistamiseksi. Riskiarvion luomisessa on otettava huomioon ilmoitusvelvollisen toiminnan luonne, koko ja laajuus. Edellä mainittujen tekijät huomioon ottaen

ilmoitusvelvollisen on luotava riittävät toimintaperiaatteet, menettelytavat ja tarvittava valvonta rahanpesun ja terrorismin rahoittamisen riskien vähentämiseksi. (Rahanpesulaki 2 luku 3§)

Ilmoitusvelvollisella viitataan Rahanpesulaki 1 luvun 2 §:n 1 momentissa määriteltyjä yhteisöjä ja elinkeinonharjoittajia. Ilmoitusvelvollisiin kuuluu esimerkiksi luottolaitokset, vakuutusyhdistykset, joukkorahoituksen välittäjät, luotonvälittäjät, tilintarkastajat ja rahoituspalveluja tarjoavat yritykset. (Rahanpesulaki 1 luku 4§)

2.1 Asiakkaan tunteminen ja tunnistaminen

Asiakkaan tuntemisella ja asiakkaan tunnistamisella tarkoitetaan sitä, että valvottava tunnistaa ja tuntee asiakkaan toiminnan laadun ja laajuuden. Valvottavat tahot ovat finanssivalvonnan valvomia tahoja, eli esimerkiksi pankit, vakuutus- ja eläkeyhtiöt, sijoituspalveluyritykset, rahastoyhtiöt ja pörssi (Finanssivalvonta). Asiakkaan tuntemiseen sisältyy menettelyt, joiden avulla voidaan varmistua asiakkaan henkilöllisyydestä, tuntee asiakkaan toiminnan ja sen taustoja asiakassuhteen edellyttämällä laajuudella. Asiakkaan tunnistaminen ja asiakkaan henkilöllisyyden varmistaminen ovat keskeinen osa tunnistamisprosessia ja sen avulla varmistetaan, kenen kanssa asioidaan. Valvottavalla lähtökohtaisesti ei saa olla anonyymejä asiakkaita ja valvottavalla on oikeus olla hyväksymättä asiakkaakseen sellaista tahoa, joka ei anna itsestään tai toiminnastaan tarvittavia tietoja. Valvottavan ei myöskään tarvitse ottaa asiakkaakseen sellaista tahoa, jonka toiminta muodostaa tavanomaista suuremman riskin rahanpesun tai terrorismin rahoittamisen näkökulmasta. (Finanssivalvonta, 2015 s.12, 18)
Rahanpesulain 3 luvun 1-3§ :n mukaan ilmoitusvelvollinen ei saa perustaa asiakassuhdetta, suorittaa liiketointa tai ylläpitää liikesuhdetta, jos ilmoitusvelvollinen ei pysty toteuttamaan rahanpesulain 3 luvun mukaan säädettyjä toimia. Jos ilmoitusvelvollinen on luottolaitos, se ei saa toteuttaa maksutapahtumia maksutilin kautta, jos ilmoitusvelvollinen ei pysty toteuttamaan säädettyjä toimia. (Rahanpesulaki 3 luku 3§)

Rahanpesulain 3 luvun 1§:n mukaan ”Ilmoitusvelvollisen on tunnistettava asiakkaansa ja todennettava tämän henkilöllisyys vakituista asiakassuhdetta perustettaessa. Lisäksi ilmoitusvelvollisen on tunnistettava asiakkaansa ja todennettava tämän henkilöllisyys, jos:

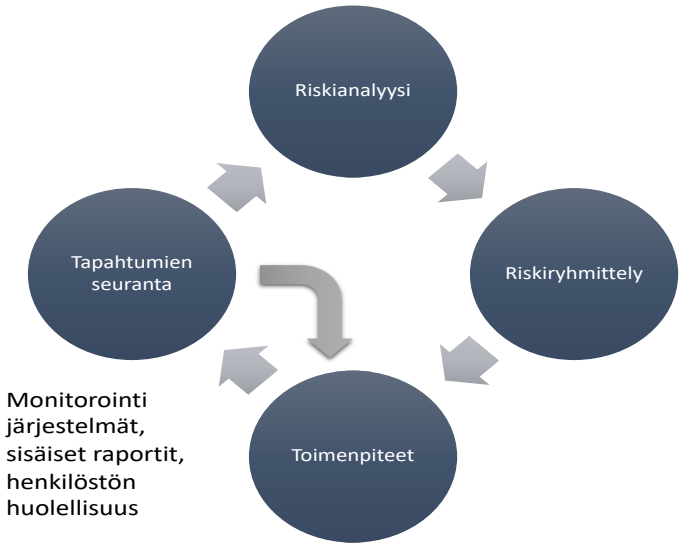
1) liiketoimen suuruus tai toisiinsa kytkeytyvien liiketoimien suuruus yhteensä on vähintään 10 000 euroa ja asiakkuus on satunnainen, tai kyse on maksajan tiedot -asetuksen 3 artiklan 9 kohdassa tarkoitetusta varojen siirrosta, jonka määrä ylittää 1 000 euroa;
2) tavaroiden myynnissä suoritettavan liiketoimen suuruus tai toisiinsa kytkeytyvien liiketoimien suuruus käteisenä on yhteensä vähintään 10 000 euroa ja asiakkuus on satunnainen;
3) kyse on epäilyttävästä liiketoimesta tai jos ilmoitusvelvollinen epäilee liiketoimeen sisältyviä varoja käytettävän terrorismin rahoittamiseen tai sen rangaistavaan yritykseen; tai
4) ilmoitusvelvollinen epäilee aiemmin todennetun asiakkaan henkilöllisyyden todentamistietojen luotettavuutta tai riittävyyttä.” (Rahanpesulaki 3 luku 2§)

Finanssivalvonta on määritellyt asiakkaan tuntemisen osa-alueet: ”

- asiakkaan ja asiakkaan edustajan tunnistaminen (identifiointi),
- asiakkaan henkilöllisyyden todentaminen (verifiointi),
- asiakkaan edustajan henkilöllisyyden todentaminen tarvittaessa,
- tosiasiallisen edunsaajan tunnistaminen ja henkilöllisyyden todentaminen tarvittaessa,
- tietojen hankkiminen liikesuhteen tarkoituksesta ja laadusta (tietojen hankkiminen asiakkuudesta),
- tietojen dokumentointi ja säilyttäminen määräajan,
- liiketoimien ja asiakassuhteiden jatkuvan seurannan järjestäminen riskiperusteisesti ja
- selonottovelvollisuuden noudattaminen.” (Finanssivalvonta 2015, s. 12-13)

Asiakkaan tunnistamisprosessissa sovelletaan riskiperusteista lähestymistapaa. Valvottavalla tulee olla riskiperusteista seurantaa varten tarvittavat riskienhallintajärjestelmät, joiden avulla se arvioi asiakkaista aiheutuvia riskejä. Riskiperusteisella lähestymistavalla tarkoitetaan sitä, että valvottava suhteuttaa tarvittavat seuranta ja tunnistamistoimet asiakkaan aiheuttaman riskin perusteella. Valvottavan tulee kohdistaa tehostettuja toimenpiteitä, jos asiakas kuuluu riskiryhmään tai jos asiakkaan liiketoimintaan liittyy normaalia enemmän rahanpesuun liittyviä riskejä. Kuvassa 2. on havainnollistettu riskiperusteinen asiakkaan arviointi ja tuntemisprosessi. Riskiperusteinen arviointi on jatkuva prosessi, jossa asiakkaalle tehdään riskianalyysi ja tämän jälkeen riskiryhmittely, joka määrittää tarvittavat toimenpiteet. Toimenpiteisiin kuuluu monitorointi järjestelmät, sisäiset raportit ja henkilöstön huolellisuus asiakkaan

tunnistamisvaiheessa. Tämän jälkeen tapahtumia seurataan ja tarvittaessa tehdään lisätoimenpiteitä, jos tapahtumista huomataan poikkeavaa.



Kuva 2. Riskiperusteinen arviointi (mukaillen Finanssivalvonta 2015, s.14)

2.2 Rahanpesun estäminen ja rahanpesun menetelmät

Rahanpesun tarkoituksena on saada rikollisesta toiminnasta saadut voitot osaksi laillista rahatalouden kiertokulkua samalla piilottaen rahojen alkuperäisen lähteen. Rahanpesu koostuu yleensä lukuisista transaktioista, joiden avulla pyritään peittämään rahan alkuperäinen lähde. Rahanpesu on isossa roolissa järjestäytyneessä rikollisuudessa, koska sen avulla saatetaan luoda kulissi, jonka tarkoituksena on samalla peittää rikollista toimintaa. (Sullivan 2015, s.5-6)

Sullivan jakaa kirjassaan rahanpesun prosessin kolmeen osaan; sijoittelu, peittely ja integrointi. Sijoittelu on prosessin ensimmäinen osa, jonka tarkoituksena on pitää rikoksesta saatu raha tai omaisuus erillään rikoksesta. Rahanpesun ensimmäisessä vaiheessa käsitellään yleisesti rahaa käteisenä, joten sen säilöminen ja kuljettaminen ja pankkiin tallettaminen isoissa määrissä on haaste. Esimerkiksi huumekaupasta saadut tuotot ovat yleisesti käteisenä ja isompien ostoksien tekeminen käteisellä herättää kysymyksiä rahojen alkuperästä, joten rahat olisi saatava talletettua pankkiin. Isojen käteismäärien tallettaminen suoraan pankkiin ei myöskään onnistu, ellei pysty selvittämään rahojen alkuperää pankille. Tyypillisin tapa on jakaa iso käteismäärä

useisiin talletuksiin useamman henkilön kesken tai siirtämällä ne maantieteellisesti toiseen paikkaan. Rahanpesun toinen vaihe on peittely. Tässä vaiheessa rahat on jo saatu rahatalouden kiertokulkuun ja tarkoituksena on peittää rahojen omistaja ja alkuperä. Peittelyssä rahanpesijän tarkoituksena on tehdä lukuisia transaktioita esimerkiksi useiden yrityksien ja toimijoiden kautta. Transaktioiden avulla rikoksesta saatujen voittojen jäljittäminen on viranomaisille vaikeampaa. Peittelyyn liittyy useasti myös ulkomaalaisia pankkeja ja toimijoita. Integrointi on rahanpesun viimeinen ja kolmas vaihe. Sen tarkoituksena on saada integroitua rikoksella saadut voitot rahatalouden kiertokulkuun. Integrointi tapahtuu esimerkiksi yrityksiin sijoittamalla, ostamalla kiinteistöjä tai ostamalla kalliita luksus tuotteita. (Sullivan 2015, s. 6-12)

Ilmoitusvelvollisen on tehtävä ilmoitus rahanpesun selvittelykeskukselle, jos he havaitsevat epäilyttävää liiketoimintaa. Epäilyttävä toiminta arvioidaan sen perusteella, mikä on normaalia kyseiselle toiminalle tai toimialalle. Epäilyttävänä toimintana voidaan myös pitää sellaista toimintaa, joka eroaa asiakkaan ennakkotietojen mukaan, on itse ilmoittanut tai, jos asiakkaan toiminta muuttuu sellaisella tavalla, että siihen ei ole järkevää selitystä. Ilmoitusvelvollisuus ei edellytä ilmoitusvelvollista arvioimaan onko tapahtunut rikosta, koska rahanpesuilmoitus ei ole rikosilmoitus. (Tarvainen 2019, s. 1-2)

Keskusrikospoliisi on määritellyt rahoitusalanedustajille yleisimmät indikaattorit, joihin tulee erityisesti kiinnittää huomiota. Indikaattorit ovat ohjenuorana, jotka auttavat tunnistamaan mahdollisia rahanpesun epäilyjä, mutta ei ole tae rahanpesusta tai rikollisesta toiminnasta. Yleisimmät indikaattorit on jaettu kahdeksaan osaan; Asiakasprofiiliin liittyvät indikaattorit, asiakkaan tililiikenne, käteisvarat, kansainväliset varainsiirrot, tiedot ja asiakirjat, oikeushenkilöt, lainat ja korruptionindikaattorit. Indikaattoreja asiakasprofiiliin liittyen on esimerkiksi: ”

- Tilitoiminta on ristiriidassa asiakkaan asiakastietojen tai asiakasprofiilin kanssa;
- Useilla henkilöillä on käyttöoikeus tiliin, mutta henkilöillä ei näytä olevan perhe- tai liikesuhdetta toisiinsa;
- Tilinomistaja ei harjoita liiketaloudellista toimintaa, mutta tiliä käytetään erilaisiin taloudellisiin liiketapahtumiin;
- Sama henkilö on avannut useita tilejä, joille tehdään lukuisia pieniä talletuksia;

- Asiakas omistaa useita eri pankkitilejä tai ulkomaalaisia tilejä ilman liiketaloudellista, juridista, verotuksellista tai kirjanpidollista perustetta;
- Asiakkaan ilmoittama ammattinimike tai palkkatulo ei ole oikeassa suhteessa liiketapahtuman tasoon tai tyyppiin, esimerkiksi opiskelija tai työtön henkilö vastaanottaa tai suorittaa suuria määriä pankkisiirtoja tai henkilö tekee päivittäin suuria käteisnostoja; …"

Tarvaisen raportissa on listattu indikaattoreja myös liittyen valuutanvaihtoon, vakuutustuotteisiin, kansainväliseen kauppaan, kasinot ja rahapelit, lakimiehiin ja vastaaviin palveluihin, kiinteistönvälittäjiin, kirjanpitoon, veroparatiiseja, bulvaaneja, virtuaalivaluuttoja ja kansalaisjärjestöjä. (Tarvainen 2019, s. 1-4, s. 9-10)

2.3 Rahanpesun estämisen haasteet ja riskit pankille

Tässä kappaleessa käsitellään rahanpesun estämiseen liittyviä haasteita ja mitä riskejä rahanpesu aiheuttaa pankeille ja muille finanssialan toimijoille.

Ilmoitusvelvollisilla, joiden asiakassuhteet ovat lyhyitä tai satunnaisia voi kattavan asiakasprofiilin luominen ja asiakkaan syvempi tunteminen ja asiakkaan liiketoiminnan seuranta olla haasteellisempaa, kuin sellaisilla ilmoitusvelvollisilla, kenen asiakassuhteet ovat pitkäkestoisia. (Tarvainen 2019, s.3)

Breslow et al. (2017) kertoo artikkelissa, että finanssisektorin rahanpesun estämisen isoimpia haasteita on huono laatuinen data, koska siinä puutteita ja dataa on useasti monesta eri lähteistä ja näiden yhdistäminen on vaikeaa. Samalla pankit joutuvat monesti käyttämään henkilötyötunteja asiakkaan puuttuvien tietojen kysymiseen ja täyttämiseen. Toisena haasteena havaittiin asiakkaiden riskiarviointi ja epäilyttävien transaktioiden havainnoiminen aiheuttaa paljon virheellisiä-positiivinen hälytyksiä, joka johtaa henkilötyötuntien hukkaamiseen turhien hälytyksien selvityksessä. Haasteeksi ilmeni myös hajautetut järjestelmät ja alustat, joka vaikeuttaa transaktioiden seuraamisen automatisointia ja asiakkaan tuntemista. Epäilyttäviä transaktioita tutkivilla työntekijöillä menee todennäköisesti iso osa ajasta tiedon keräämiseen, eikä tutkivaan työhön, joka johtuu osaltaan hajautetuista järjestelmistä. (Breslow et al. 2017)

Reese tuo kirjoituksessaan esille tietosuoja-asetuksen ja rahanpesun estämisen yhdistämisen haasteita. EU:n tietosuoja-asetus rajoittaa asiakkaasta tallennettavaa tietoa ja sen käyttöä, mutta samalla rahapesun estämistä varten asiakkaasta pitää kerätä ja tallentaa tietoa ja näiden lakien ristiriita saattaa aiheuttaa haasteita pankille. Tietosuoja-asetus määrää mitä henkilötietoja voidaan milloinkin kerätä ja pankkien on tuotava esille, miksi niitä kerätään ja mihin tarkoitukseen. Samalla rahanpesun estämisen lainsäädäntö vaatii pankkia keräämään asiakkaasta tietoa ja prosessoimaan sitä ja arvioimaan asiakkaan riskiä. Suuri osa rahanpesun lainsäädännön nojalla asiakkaasta kerätyt tiedot kuuluvat osaksi tietosuoja-asetusta. Pankkien on siis tärkeä varmistaa, että rahanpesulainsäädännön nojalla kerätyt tiedot eivät riko tietosuoja-asetuksen sääntöjä. (Reese 2018)

Pankin ollessa osallisena rahanpesuun se saattaa korvauksien maksamisen lisäksi kärsiä imagohaitasta ja pörssikurssin laskusta. Esimerkiksi tanskalainen pankki Danske Bank ja tarkemmin Danske Bankin Viron-yksikkö on epäiltynä osallistumisesta rahanpesuun. Tämä voi johtaa miljardiluokan korvauksiin ja Brännären artikkelin julkaisu hetkellä lokakuussa 2018 yrityksen pörssikurssi oli romahtanut 34 prosenttia vuoden 2018 alusta laskettuna. (Brännare 2018)

Pankille voi koitua pienemmästäkin rikkeestä mainehaittaa, koska pankki voi joutua viranomaisten ylläpitämälle julkiselle mustalle listalle. Musta lista on aluehallintoviraston ylläpitämä sivu, johon yritykset joutuvat laiminlyödessään ilmoitusvelvollisuuttaan. Poliisiammattikoulun tutkijan Pirjo Jukaraisen mielestä ”Yritys, joka laiminlyö asiakkaiden seurannan ja asiakkaiden riskiperusteisen arvioinnin ja saa siitä julkisen mainehaitan, se on vieläkin tehokkaampi keino kuin yksittäisen rahanpesijän saama tuomio”. Julkiselle listalle voi päätyä kuka tahansa ilmoitusvelvollinen, kuka ei ole huolehtinut riskien arvioinnista omassa toiminnassaan. (Ikävalko 2019)

Ilmoitusvelvollisen rikkoessa rahanpesulaissa määrättyjä vastuita asiakkaan tuntemisen ja tunnistamisen vaatimuksia, tai ei toteuta tarpeellista riskiarviota voi siinä tapauksessa Finanssivalvonta määrätä ilmoitusvelvolliselle rikemaksun. Rikemaksun suuruus määräytyy rikkeen laatu, laajuus ja kestoaika. Rikemaksu on oikeushenkilölle vähintään 5000 euroa ja enintään 100 000 euroa ja luonnolliselle henkilölle vähintään 500 euroa ja enintään 10 000 euroa. Finanssivalvonta voi määrätä ilmoitusvelvolliselle myös seuraamusmaksun, jos

huolimattomuus on vakavaa, toistuvasti tai tahallaan laiminlyö ilmoitusvelvollisuuden rahanpesulaissa määrättyjä vastuista. Luotto- ja rahoituslaitoksille seuraamusmaksun suuruus on enintään kymmenen prosenttia luotto- tai rahoituslaitoksen edeltävän vuoden liikevaihdosta tai viisi miljoonaa euroa, sen mukaan kumpi on suurempi. Seuraamusmaksu saa kuitenkin olla enintään kaksi kertaa isompi, kuin laiminlyönnillä saatu hyöty, jos hyödyn määrä on määriteltävissä. (Rahanpesulaki 8 luku 1-4§)

Esimerkkinä Finanssivalvonnan seuraamusmaksusta on Finanssivalvonnan 18.12.2019 julkaiseman tiedotteen mukaan heidän määräämänsä 980 000 euron seuraamusmaksu S-Pankki Oy:lle. Finanssivalvonta määräsi S-Pankille seuraamusmaksun, koska S-Pankki ei ollut noudattanut tarpeeksi kattavaa riskiperusteista toimintatapaa, eikä ollut hankkinut asiakkailta riittäviä tuntemistietoja. Finanssivalvonta antoi samassa tiedotteessa julkisen varoituksen FIM Varainhoito Oy:lle, koska heillä oli puutteita asiakkailta hankituissa tuntemistiedoissa. Laiminlyönnit ilmenivat Finanssivalvonnan vuosina 2017 ja 2018 tekemän tarkastuksen aikana. Finanssivalvonta ei kuitenkaan epäile, että S-Pankki Oy tai FIM Varainhoito Oy olisi syyllistynyt rahanpesurikoksiin. (Finanssivalvonta 2019)

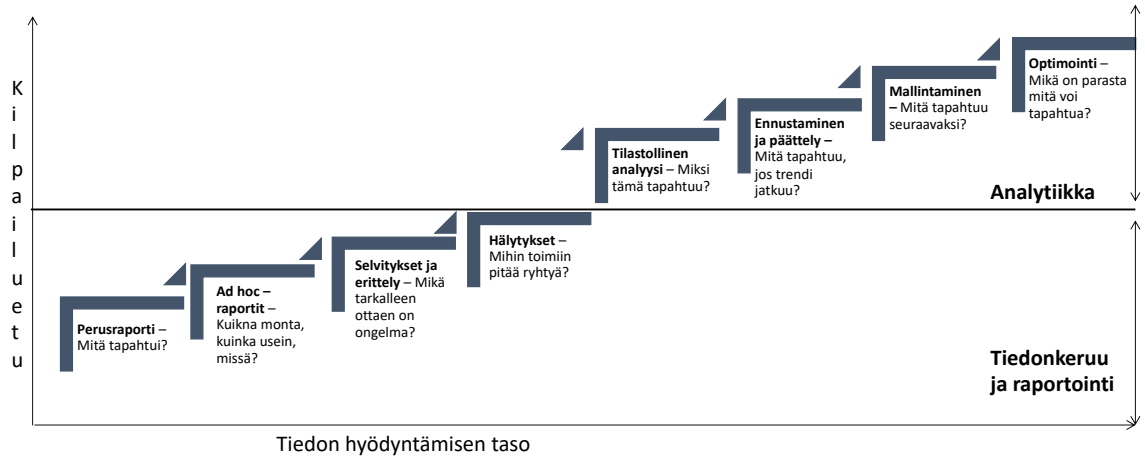
3 DATA-ANALYTIIKKA

Tässä luvussa käsitellään alkuun data-analytiikalle merkittävät perusteet, eli analytiikan ja datan merkitys. Tämän jälkeen siirrytään käsittelemään data-analytiikan prosesseja. Prosessien jälkeen siirrytään big dataan ja koneoppimiseen. Näiden ydinasioiden jälkeen perehdytään data-analytiikan haasteisiin ja erityisesti datan laatuun. Viimeisenä käsitellään data-analytiikan hyödyt ja mitä lisäarvoa organisaatio voi saada data-analytiikan hyödyntämisestä.

Analytiikalla tarkoitetaan datan kattavaa hyväksikäyttöä, kvantitatiivista ja tilastollista analyysiä. SAS määrittelee analytiikan datan ja matematiikan yhdistämiseksi, jonka avulla voidaan vastata liiketoiminnan kysymyksiin, ennustaa lopputuloksia ja automatisoida päätöksentekoa (SAS 2019a). Analytiikkaa hyödynnetään selittävissä ja ennustavissa malleissa ja näitä analyysejä voidaan hyödyntää toiminnan, päätöksenteon ja johtamisen tukena. Analytiikka on osa business inteligenceä, eli älykästä tiedonhallintaa. (Davenport & Harris 2007, s. 26)

Kuva 3. havainnollistaa business inteligenceä ja analytiikan osuutta siitä. Business inteligence on lähellä analytiikkaa, mutta se on keskittynyt liiketoimintaan, etsimään trendejä ja on enemmän kuvailevaa analytiikkaa kuin ennustavaa analytiikkaa (Dataquest 2019). Kuva on jaettu kahteen osaan, tiedonkeruu ja raportointi ja analytiikkaan. Kuvan vaaka-akselilla havainnollistetaan tiedon hyödyntämisen tasoa, joka kasvaa oikealla mentäessä. Pystyakselilla havainnollistetaan prosessin tuottamaa mahdollista kilpailuetua, joka kasvaa ylöspäin mentäessä. Tiedonkeruu ja raportointi on perinteinen datan hyödyntämisen osa-alue, jossa ei varsinaisesti hyödynnetä analytiikkaa. Tässä kategoriassa prosesseissa hyödynnetään olemassa olevaa dataa, ja tehdään sen avulla raportteja ja esimerkiksi perusraportoinnissa voidaan datan avulla selvittää, mitä tapahtui. Ad hoc-raportit ovat syvällisempiä raportteja, kuin perusraportit ja hyödyntävät enemmän saatavilla olevaa tietoa. Ad hoc-raporteille pystytään lisäämään perusraporttiin esimerkiksi tietoa missä tapahtui tai kuinka monta kertaa tapahtui. Selvitykset ja erittely prosessissa dataa hyödynnetään jo enemmän ja pystytään vastaamaan, mikä on tarkalleen ongelma. Hälytykset prosessissa pystytään luomaan ilmoituksia tapahtumista ja kuinka niihin pitäisi reagoida. (Davenport & Harris 2007, s. 27)

Kuva 3. ylälaidassa prosesseissa siirrytään hyödyntämään dataa tehokkaammin ja soveltamaan siihen analytiikkaa. Ensimmäisenä prosessina on tilastollinen analyysi, jonka avulla pystytään vastaamaan, miksi jokin asia tapahtui. Ennustamisessa ja päättelyssä tarkoituksena on katsoa tulevaisuuteen ja dataan pohjautuvan analytiikan avulla selvittää, esimerkiksi mitä tapahtuu, jos nykyinen trendi jatkuu. Mallintamisessa nimensä mukaan mallinnetaan tulevaa ja pystytään vastaamaan mitä tapahtuu seuraavaksi. Optimointi on prosesseista se vaihe, jossa hyödynnetään saatavilla olevaa dataa mahdollisimman tehokkaasti ja sen avulla pystytään vastaamaan, mikä on parasta mitä voi tapahtua. Optimoinnin avulla saavutetaan näistä prosesseista myös suurin kilpailuetu. (Davenport & Harris 2007, s. 27)



Kuva 3. Tiedon hyödyntäminen (Davenport & Harris 2007, s. 27).

Data-analytiikalla tarkoitetaan prosessia, jossa tutkitaan saatavissa olevaa dataa ja pyritään tunnistamaan datasta haluttua informaatiota käyttäen hyväksi siihen saatavilla olevia menetelmiä, ohjelmistoja ja järjestelmiä. (Rouse 2016)

Data-analytiikka on prosessi, jossa tutkitaan dataa tiettyjen toimintatapojen ja ohjelmistojen avulla. Data-analytiikalla pyritään tunnistamaan datan sisältämää informaatiota. Data-analytiikalla prosessoidaan dataa ja pyritään etsimään siitä haluttuja tuloksia. Saatuja tuloksia pystyy hyödyntämään hyvin laajasti yrityksen liiketoiminnassa ja erinäisillä tieteenaloilla auttamaan päätöksenteossa, teorioiden, mallien tai hypoteesien todistamisessa. Data-analytiikan avulla voidaan löytää sellaista informaatiota, joka normaalisti pelkästä datasta ei tulisi esille (Frankendfield 2019). Data-analytiikkaa voidaan hyödyntää myös kasvattamaan yrityksen myyntiä, kehittämään markkinointia ja esimerkiksi reagoimaan markkinoilla tapahtuviin muutoksiin nopeammin. (Rouse 2016)

Terminä data-analytiikka ensisijaisesti viittaa laajaan käyttökohteiden kokoelmaan. Data-analytiikalla voidaan viitata perinteiseen business inteligenceen, raportointi työkaluihin, verkossa toimiviin analytiikka työkaluihin ja moniin kehittyneen analytiikan muotoihin. Tämän perusteella data-analytiikka on hyvin lähellä business analytiikkaa, mutta sillä erolla, että business analytiikka on keskittynyt yrityksen liiketoimintaan ja data-analytiikalla on laajempi käyttötarkoitus. (Rouse 2016)

Data-analytiikka voidaan myös jakaa kvalitatiiviseen ja kvantitatiiviseen data-analytiikkaan. Kvalitatiivisessa data-analytiikassa analysoidaan dataa, joka on numeerisessa muodossa, jota pystyy vertailemaan tai mittaamaan tilastollisesti. Kvantitatiivisessa data-analytiikassa keskitytään dataan, joka on muissa kuin numeerisessa muodossa, eli esimerkiksi tekstinä, kuvina, videona tai äänenä. Data-analytiikan avulla voidaan esimerkiksi kerätä suuria määriä dataa ja testata, voidaanko analysoidulla datalla todistaa tai tukea tutkittavana olevaa asiaa (Early 2015, s.495). (Rouse 2016)

3.1 Data

Data on tosiasioihin perustuvaa informaatiota, mutta sitä ei ole käsitelty millään tavalla. Käsittelemättömästä datasta voidaan käyttää myös termiä raakadata. Raakadataa on saatavilla monesta eri lähteestä, esimerkiksi tietojärjestelmistä tai sensoreista ja data voi olla samalla myös monessa eri muodossa. Raakadata ei itsessään tuo lisäarvoa, ennen kuin se on käsitelty. Datasta saadaan tehtyä informaatiota, kun se käsitellään yleisesti ymmärrettävään muotoon. Informaatio sisältää dataa, mutta data ei välttämättä sisällä informaatiota ja datan sisältämää informaatiota ei saa selville, ennen sen käsittelyä. Datasta muodostuu tietämystä, kun henkilöllä on kyky tulkita sen sisältämää informaatiota ja tehdä sen pohjalta oikeita johtopäätöksiä. Datasta saadusta tietämyksestä kehittyy viisautta, kun henkilö kykenee yhdistämään tietoa eri lähteistä ja havainnollistamaan vaihtoehtoiset toimintatavat aiemmin saadun tietämyksen pohjalta, sekä vertailemaan olemassa olevia vaihtoehtoja. (Ahsan & Shah 2019; Ahonen et. al 2017 s. 19-20)

Dataa on monen tyyppistä ja ne voidaan lähtökohtaisesti jakaa neljään kategoriaan; nominaalisuus, ordinaalisuus, intervallisuus ja suhteellisuuteen (Devi & Murty 2015 s. 41). Datan lajittelu selkeisiin kategorioihin mahdollistaa systemaattisen mittaamisen ja analysoinnin datalle, joka ei normaalisti olisi laskettavissa, nämä mittausyksiköt ovat tärkeitä, koska kun tiedetään mihin näistä data kuuluu, osataan valita oikeat tekniikat datan käsittelyyn (Gupta 2016 s.11)

Nominaaliasteikko on kuvaileva asteikko, eli se luokittelee datan esimerkiksi toimialoihin tai onko maa-alue esimerkiksi mäkistä, metsää tai järveä. Nominaaliasteikossa dataa pystyy vertaamaan vaan yhtäläisyyksien perusteella, eli onko toisella datalla sama nominaalisuus vai

eri. Nominaaliasteikossa olevalla datalla ei ole paremmuus tai suuruus eroa toiseen. Nominaaliasteikossa olevalle datalle pystyy antamaan numeerisen termin, eli voidaan antaa automerkille numero 1 ja toiselle automerkille numero 2. Tämä ei kuitenkaan tarkoita, että toinen automerkki olisi kaksi kertaa enemmän jotain kuin ensimmäinen automerkki, eikä numeroilla voi verrata korkeampaa asemaan toiseen, kuten esimerkiksi kilpaurheilussa sija 1 on parempi kuin sija 2. Nominaaliasteikossa numeeriset arvot ovat vain kategorisoinnin apuna. (Gupta 2016 s.11-12)

Nominaaliasteikossa data voi olla binäärisiä tai ei-binäärisiä. Binäärisessä nominaaliasteikossa datan arvolla on vain kaksi vaihtoehtoa, esimerkiksi juomavaihtoehtoja on tee ja kahvi. Ei-binäärisessä asteikossa datalla on useampi kuin kaksi vaihtoehtoa, esimerkiksi TV-valmistajia on enemmän kuin kaksi. (Devi & Murty 2015 s. 41-42)

Ordinaaliasteikossa ilmaistaan datan järjestystä, mutta ei eroja niiden välillä eli esimerkiksi alhainen, keskiverto ja korkea. Ordinaaliasteikkoa käytetään datan kategorisoimiseen. Sen avulla pystytään vertailemaan dataa ilman, että tarvitaan tietoa kategorioiden erojen suuruksista. Ordinaaliasteikkoa pystyy käyttämään esimerkiksi mielipidekysymyksien tuloksissa tai maanjäristysten voimakkuuksissa. Nominaaliasteikossa olevan datan pystyy muuntamaan ordinaaliasteikkoon, jos datan pystyy muokkaamaan järjestelmälliseen muotoon, esimerkiksi nominaaliasteikkossa olevat värit voitaisiin muokata ordinaaliasteikkoon käyttämällä värien aallonpituuksia ja järjestämään ne sen mukaan (Devi & Murty 2015 s. 46). Ordinaaliasteikolla olevasta datasta ei kuitenkaan voi tehdä laskennallisia päätöksiä, koska esimerkiksi ordinaaliasteikolla voidaan verrata veden puhtautta. Jos veden puhtaus määritellään asteikolla yhdestä viiteen, ykkösen ollessa puhtain mahdollinen arvo, puhtausarvon kaksi saanut vesi ei välttämättä ole kaksi kertaa puhtaampaa kuin puhtausarvon neljä saanut vesi. Ordinaaliasteikolla voidaan datasta erottaa suhteellisia eroja, joten siihen ei voi käyttää keskiarvoa, mutta siihen voi soveltaa esimerkiksi mediaania tai moodia. Devi & Murty (2015, s.46) tuovat esille, että ordinaaliasteikossa olevalle datalle voi soveltaa persentiiliä, jos datan arvot ovat järjestyksessä.(Gupta 2016, s.11-13)

Intervalliasteikossa pystytään ilmaisemaan datan järjestystä samalla tavalla, kuin ordinaaliasteikossa, mutta intervalliasteikossa tiedetään tarkasti havaintojen välimatkat. Esimerkiksi ulkolämpötila on 40 °C on suurempi kuin 0 °C ulkolämpötila. Intervalli asteikossa

on tärkeä ottaa huomioon, että sen dataan ei voi soveltaa yhteen tai jakolaskuja, mutta sen datasta voidaan laskea eroja. Intervalliasteikko voi määrittää esimerkiksi lämpötilaa, koordinaattisijaintia tai päivämäärää. (Gupta 2016, s.11-14)

Suhdelukuasteikossa data on hyvin samanlaisessa muodossa, kuin intervalliasteikossa, mutta datalla on selkeä nolla-arvo. Esimerkiksi, kun asukastiheys on nolla, kyseisellä alueella ei ole ollenkaan asukkaita, eikä asukastiheys voi olla pienempää, kuin nolla. Ikä on esimerkiksi määre, joka kuuluu osaksi suhdelukuasteikkoa. Siinä on selkeä nolla-arvo, joku ei voi olla alle 0 vuotiasta, mutta pystytään sanomaan, että 10-vuotias on kaksi kertaa vanhempi kuin 5-vuotias. Voidaan myös verrata, että henkilöillä, jotka ovat 6 ja 8-vuotiata on sama ikäero kuin 10 ja 12-vuotiaalla. Suhdelukuasteikko on suosituin lukuasteikko datan louhinnassa, kuvion tunnistuksessa ja koneoppimisessa sen monipuolisten käyttömahdollisuuksien ansiosta(Devi & Murty 2015 s. 49). Nominaalinen ja ordinaalinen data voidaan luokitella kategorisoitavaksi dataksi ja intervallinen ja suhdelukuinen data voidaan luokitella numeeriseksi dataksi. (Gupta 2016, s. 14)

3.2 Data-analytiikan prosessi

Data-analytiikan prosessin tarkoituksen on saada tuotettua saatavilla olevasta raakadatasta informaatiota, tietämystä, viisautta ja lopuksi auttaa päätöksenteossa. Datan analysointimenetelmät riippuvat sen muodosta. Analysointimenetelmät on valittava tapauskohtaisesti ja ensisijaisesti sen mukaan onko data numeerisessa vai tekstimuodossa. Alla olevassa Kuvassa 4. on havainnollistettu yleisellä tasolla datan jalostamista päätöksentekoon. Ensimmäisessä vaiheessa on itse raakadatan kerääminen, jota tässä prosessissa lähdetään jalostamaan. Vaiheessa kaksi siirrytään käsittelemään dataa siten, että siitä saataisiin ulos sen sisältämää informaatiota ja tässä vaiheessa voidaan tuottaa jo kuvailevaa data-analyysiä ja visualisoimaan aiempaa raakadataa. Kolmannessa vaiheessa hyödynnetään aiemmin saatua informaatiota aineiston mallintamisessa, kvantitaavisen ja kvalitatiivisen tiedon yhdistämisessä. Neljännessä vaiheessa on yhdistetty kolmannen vaiheen tietämys viisaudeksi ja silloin on saatavilla päätösvaihtoehdot ja niiden vertailu. Lopputuloksena on viimeinen eli viidesvaihe, jossa tehdään lopullinen päätöksenteko. Prosessin viimeisissä vaiheissa on tärkeä ottaa huomioon se, että pelkkä aineiston kerääminen ja käsittely ei riitä kattavan kuvan

saamiseksi, koska tietämys ja viisaus vaativat laajempaa ymmärrystä käsiteltävästä kohteesta. Kerätyn aineiston lisäksi on hyvä ottaa huomioon liiketoimintaympäristöä kuvaavaa tietoa, sekä asiantuntijoilla olevaa hiljaista tietoa. (Ahonen et al. 2017, s. 20-21)

Kuva 4. Datan analysointiprosessi (Ahonen et al. 2017, s. 20).

Raakadatan keräämiseen voidaan käyttää datan louhinnaksi kutsuttua prosessia. Datan louhinta on terminä yleistävä ja sisältää kattavan kirjon metodeja, algoritmeja ja teknologioita. Yhdistävä tekijä näillä on se, että niitä hyödynnetään raakadatan keräämiseen isoista tietomassoista, eli esimerkiksi tietokannoista. (Xanthopoulos et al. 2013, s. 1)

Datan louhintaa voidaan käyttää hyväksi esimerkiksi, kun pyritään tunnistamaan tilisiirroista tai luottokortin käytöstä haitallista tai laitonta toimintaa. Data louhinta toimii myös monessa muussa, esimerkiksi kun haetaan isoja määriä dataa verkkosivuilta tai teollisuuden sensoreista. Ihmisten käyttäytymistä voidaan myös analysoida data louhinnan avulla. Raakadata on useasti hyvin monessa eri muodossa ja se syötetään data louhinnan prosessien läpi, jonka jälkeen raakadata on käytettävässä muodossa. Data louhinnan prosessissa data kerätään, siivotaan ja muokataan standardoituun muotoon. Tämän prosessin jälkeen se voidaan prosessoida analyyttisesti ja saada siitä informaatiota. (Charu 2015, s. 1-4)

3.3 Big data

Big dataksi voidaan kutsua dataa, jonka määrä, monimuotoisuus ja kasvunopeus on niin suurta, että sen käsittely normaalien tietokantojen ja työkalujen avulla on vaikeaa. Datan määrän kasvaessa suureksi, sen tallentaminen, prosessointi ja analysoiminen on entistä vaikeampaa ja siihen vaaditaan big datan käsittelyyn suunniteltuja työkaluja. Big datassa uutta dataa syntyy nopeassa tahdissa ja useasti sitä halutaan analysoimaan reaaliajassa, joka vaatii omat työkalunsa. Big datalle on myös ominaista, että se sisältämä data voi olla monessa eri muodossa (Exasol 2019). (Bahga & Madisetti 2016, s. 25)

Viime vuosina datan määrä on kasvanut eksponentiaalisesti ja esimerkiksi Twitterin käyttäjät lähettävät minuutissa noin 300 000 twiittiä ja YouTubeen lisätään minuutissa noin 300 tunnin edestä videoita ja Amazon sivuilla vierailee minuutissa noin 4300 ihmistä. Big dataa syntyy myös esimerkiksi teollisuuden sensoreista, terveydenhuollon laitteistojen tiedoista, pankin asiakkaiden transaktioista, käyttäjien toiminnasta verkkokaupoissa ja sosiaalisen median postauksista. Kaikesta tästä toiminnasta tuotetaan dataa ja sen määrä on tulevaisuudessa vain kasvussa. Big datan kasvun ansiosta sitä voidaan hyödyntämään laaja-alaisesti esimerkiksi yrityksien markkinoinnin ja myynnin tehostamisessa, finanssialalla, teollisuudessa ja terveydenhuollossa. (Bahga & Madisetti 2016, s. 25-26)

Big datan määritelmä on lähtöisin vuodelta 1997 Gartnerin kolmen V:n mallista, jossa big data määriteltiin kolmen piirteen avulla, jotka olivat määrä (volume), nopeus (velocity) ja monimuotoisuus(variety). Tämän jälkeen big datan määritelmään on lisätty useita piirteitä. (Buyya et al. 2016, s. 7-8)

Big data määritellään nykypäivänä useasti viiden V:n avulla eli viiden piirteen avulla, jotka ovat määrä (volume), nopeus (velocity) monimuotoisuus (variety) ja epävarmuus (veracity) ja arvo (value) (Bahga & Madisetti, 2016 s. 26). Microsoft on kehittänyt itse kuuden V:n mallin big datalle, jossa viiden V:n malliin on lisätty näkyvyys (visibility). Microsoftin kuuden V:n mallissa arvo piirteen tilalle on otettu vaihtelevuus (variability). (Buyya et al. 2016, s. 8)

Viiden V:n ja Kuuden V:n mallissa määrä kuvaa sitä, että dataa on niin paljon tallennettavaksi, että sen tallentaminen yhdelle tietokoneelle ei ole kannattavaa, vaan se täytyy jakaa useammalle

tietokoneelle. Esimerkiksi sosiaalisen median alustat prosessoivat päivässä miljardeja viestejä tai teollisuudessa syntyy usean teratavun edestä dataa päivässä pelkästään sensoreista. Datan määrä on koko ajan kuitenkin myös kasvussa, joten sen tallentamiseen ja prosessointiin vaaditaan erityisiä työkaluja ja ratkaisuja. Big datan toinen piirre on nopeus. Nopeudella viitataan siihen, kuinka nopeasti uutta dataa syntyy ja se on yksi pääsyistä siihen, minkä takia datan määrä kasvaa tällä hetkellä eksponentiaalisesti. Datan nopea syntyminen johtaa siihen, että tallennetun datan määrä kasvaa isoksi hyvin lyhyessä ajassa. Haasteita nopeasta datan syntymisestä tulee myös silloin, kun dataa pitää pystyä käsittelemään reaaliajassa, esimerkiksi tehtaiden sensoreiden tuottamaan dataa. Kolmas big datan piirre on monimuotoisuus ja sillä viitataan siihen, että big data järjestelmien data voi olla strukturoitua, strukturoimatonta tai puoliksi strukturoitua dataa ja sisältää tekstiä, kuvia, videoita tai dataa sensoreista. Neljäs piirre on epävarmuus. Epävarmuudella viitataan siihen, että kuinka hyödyllistä data on. Dataa kertyy hyvin paljon ja se saattaa sisältää paljon kohinaa ja se joudutaan poistamaan ennen kuin siitä on mahdollista saada merkitsevää ja tarkkaa dataa. Viides piirre on arvo. Arvolla viitataan siihen, kuinka hyödyllistä data on sen käyttökohdetta ajatellen. Jokaisen big data analytiikka järjestelmän tavoitteena on saada tuotettua datasta lisäarvoa. Lisäarvon määrää voidaan myös mitata sen perusteella, kuinka nopeasti kyseinen järjestelmä pystyy käsittelemään dataa. Kuuden V:n mallissa Microsoft korostaa näkyvyydellä kokonaiskuvan hahmottamisen tärkeyttä, ennen kuin datasta voidaan tehdä päätöksiä. Vaihtelevuudella Microsoft tarkoittaa sitä, kuinka monimuotoista data on ja kuinka monessa eri muodossa data on. (Bahga & Madisetti 2016, s. 26-27; Buyya et al. 2016, s. 8)

Big data ratkaisuja pystyy käyttämään esimerkiksi pankki ja finanssisektorilla petosten paljastamiseen. Big datan avulla pystytään estämään erityisesti luottokortti huijauksia, rahanpesua ja vakuutus huijauksia. Big data ratkaisujen avulla pystytään analysoimaan dataa monesta lähteestä samanaikaisesti reaaliajassa ja tarkastelemaan asiakkaan transaktioita. Tämän lisäksi voidaan kehitellä koneoppimis algoritmeja tunnistamaan poikkeamia transaktioissa ja hälyttämään mahdollisesti epärehellisistä toimista. Big data ratkaisujen avulla pystyy myös analysoimaan isoja määriä transaktioita myös historiasta ja etsimään viitteitä epärehellisestä toiminnasta. (Bahga & Madisetti 2016, s. 29)

3.4 Koneoppiminen

Koneoppiminen ei ole terminä uusi, vaan siihen liittyviä algoritmeja on ollut jo 1970-luvulta lähtien. Tietokoneiden suorituskyvyn kasvaessa koneoppiminen on kasvattanut suosiotaan, koska on ollut mahdollista ratkaista koneoppimisen avulla entistä haastavampia ongelmia. Saatavilla olevan datan määrän kasvaessa on avautunut uusia kohteita koneoppimiselle. Koneoppimista käytetään tällä hetkellä esimerkiksi kasvojen tunnistukseen kuvista, big datan käsittelyyn ja markkinointiin. Koneoppimisessa tietokone oppii tekemään tietyn tehtävän, kun sille annetaan tarpeeksi dataa ja esimerkkejä tästä tehtävästä. Oppimisen jälkeen tietokone pystyy suoriutumaan itsenäisesti tästä tehtävästä uudella datalla, jota se ei ole ennen nähnyt. (Louridas & Ebert 2016, s. 110)

Bonnin (2017, s. 9) jakaa koneoppimisen oppimisprosessin kolmeen osaan; ohjattuun oppimiseen, vahvistettuun oppimiseen ja ohjaamattomaan oppimiseen. Koneoppimisen oppimisprosessin voi jakaa kahteen osaan; ohjattuun ja ohjaamattomaan oppimiseen (Louridas & Ebert 2016, s. 113). Ohjatussa oppimisessa tietokoneelle annetaan oikeaa dataa syötteeksi ja tietokoneelle annetaan samalla myös oikeat tulokset, jotka tästä datasta tulisi saada, kun siihen soveltaa tarpeelliset toimet. Tämän jälkeen tietokone päättelee tarpeelliset toimet, jotka se on tehtävä datalle päästäkseen haluttuun lopputulokseen. Ohjatussa oppimisessa tietokone hyödyntää luokittelualgoritmeja datan luokitteluun. Tietokoneelle voidaan esimerkiksi syöttää tiedot lainahakemuksista ja tiedot mitkä lainahakemuksiin liittyneet lainat jäivät asiakkaalta maksamatta. Tämän tiedon jälkeen tietokoneelle voidaan syöttää uusia lainahakemuksia ja se luokittelee lainat sen perusteella, onko todennäköisempää, että lainanhakija pystyy maksamaan lainan takaisin vai ei. Ohjaamaton oppiminen eroaa ohjatusta oppimisella sillä tavalla, että tietokoneelle ei anneta haluttua lopputulosta, vaan pelkästään lähtödata ja tietokoneen pitää sen perusteella pystyä löytämään vastaukset. Ohjaamaton oppiminen sisältää klusterointi algoritmeja, jotka analysoivat dataa ja lajittelee sen joukoiksi yhtenevien tekijöiden perusteella. Data voidaan esimerkiksi visualisoida sen sisältämien parametrien perusteella ja toisiaan lähellä olevat datapisteet kuuluvat todennäköisesti samaan kategoriaan. Vahvistetussa oppimisessa ei ole saatavilla valmiita vastauksia, vaan siinä on tarkoituksena asettaa tietokone sellaiseen ympäristöön, jossa se voi kokeilla tiettyjä toimia ja se saa niistä palautetta. Palaute perustuu siihen, tekikö tietokone halutun asian vai ei. Jokaisella yrityksellä tietokone pyrkii

maksimoimaan positiivisen palautteen määrän. (Louridas & Ebert 2016, s. 110-114; Bonnin 2017, s. 9-12)

3.5 Data-analytiikan haasteet ja datan laatu

Tässä kappaleessa käsitellään data-analytiikan haasteita ja syvennytään erityisesti datan laatuun. Ensimmäisenä käsitellään datan laatua ja sen merkitystä.

Data-analytiikka antaa yrityksille paljon, mutta siinä on myös omat haasteensa ja on hyvä ottaa huomioon. Yksi data-analytiikan haasteista liittyy dataan ja sen laatuun. Huonolaatuisen datan seuraukset voi havaita hyvin arkipäiväisessä toiminnassa. Esimerkiksi myöhästyneen kirjeen syyksi voidaan useasti postissa sanoa laitteiston vikaantumista, vaikka vika oikeasti saattoi johtua dataan liittyvästä asiasta. Kirjeen osoite saattoi olla eri, mikä on osoitetietokantaan kirjattuna. Vaihtoehtoisesti tietokantaan voi olla henkilön yhteystiedot kirjattu kahdesti, joten hänelle saattaa lähteä automaattisesti generoitavaa postia kahdesti, esimerkiksi mainoksia. (Batini & Scannapieca 2006, s. 1-2, s. 22)

Data-analytiikan haasteisiin liittyy sen vaatima osaaminen. Watson esittelee artikkelissaan analytiikan hyödyntämiseen tarvittavia taitoja. Tarvittava osaaminen on jaettu kolmeen kategoriaan; liiketoiminta osaaminen, data ja mallintaminen. Tarvittava osaaminen on niin laajaa, että kattavaa osaamista kaikkeen on vaikea saada. Liiketoiminnan osaajat ymmärtävät yrityksen liiketoimintaa ja siitä syntyvää dataa, mutta heiltä puuttuu useasti mallintamisen osaaminen. Liiketoiminta-analyytikoiden tarkoituksena on tuottaa informaatiota yritykselle analysoimalla dataa ja heillä on yleisesti ottaen osaamista liiketoiminnasta, datasta ja mallintamisesta. Data tieteiden osaajalla on kattava osaaminen datan käsittelystä ja mallintamisesta ja osaavat tehdä kattavia analyysejä datan pohjalta. Data tieteiden osaajilla on kuitenkin yleensä rajallinen liiketoimintaosaaminen. Puutteellisen osaamisen vuoksi yritykset tarvitsevat kehittyneen analytiikan tekemistä varten eri taustalta olevia työntekijöitä. Työntekijöiden osaaminen tukee toisten puutteita ja saadaan osaamista tasaisesti kaikilta osa-alueilta. (Watson 2012, s. 4-5)

Big dataan liittyviä haasteita on esimerkiksi henkilöiden yksityisyydensuojaan liittyviä uhkia. Datan määrän kasvaessa verkossa toimivat palveluntarjoajat ovat keränneet paljon

henkilötietoa ja niiden omistajuus on keskittynyt vain harvoille. Tästä voi seurata esimerkiksi liian tarkkaa käyttäjien seuraamista ja profilointia, joka voi johtaa esimerkiksi tuotteiden korkeampaan hinnoitteluun pelkästään henkilön tietojen ja profiilin perusteella. Tietoturvan kannalta kasvava datan määrä tuo itsessään myös haasteita. Valtavien tietomassojen siirtäminen ei aina ole mahdollista käsittelyä ja analysointia varten, vaan ne on tehtävä suoraan tiedon tallennuspaikassa. Tämä johtaa siihen, että sen analysointia varten on päästä suoraan käsiksi tiedon tallennuspaikkaan. Suora pääsy tietokantaan on aina tietoturvariski varsinkin, jos pääsy pitää antaa ulkopuoliselle taholle. (Rastas & Asp 2014 , s.10-11)

Datan keräämisestä on tullut helpompaa ja sen säilyttämisestä on tullut halvempaa, joka kannustaa yrityksiä tallentamaan enemmän dataa. Dataa tallentaessa on kuitenkin otettava huomioon tietosuojakysymykset, jos data sisältää henkilötietoja. Datan säilytyksen kanssa on oltava hyvin tarkka ja yrityksen on aina noudatettava EU:n tietosuoja-asetusta. Asiakkaalle tulee aina kertoa mitä tietoa kerätään ja mihin tarkoitukseen. Asiakkaalta tulee saada esimerkiksi suostumus, jos henkilötietoja käytetään sähköisessä markkinoinnissa. Tämä aiheuttaa haasteita yritykselle, koska yrityksellä täytyy olla tietosuoja-asetuksen perusteella tarkka dokumentaatio tietosuojan ja tietoturvan varmistamiseksi. Dokumentaatio pitää sisällään, kuinka yritys kerää henkilötietoja, kuinka niitä varastoidaan ja kuinka niitä käsitellään. (Markkula & Syväniemi 2015, s. 63-66)

Datan laadun dimensiot voidaan yleisesti jakaa täsmällisyyteen (accuracy), johdonmukaisuuteen (consistency) ,täydellisyyteen (completeness) , oikea-aikaisuuteen (timeliness/currency). Datan täsmällisyydellä viitataan siihen, kuinka lähellä arvo v on arvoa v'. Arvo v' on määritelty absoluuttisesti oikeaksi vastaukseksi ja arvon v:n on tarkoitus olla sama, kuin v'. Esimerkiksi henkilönnimi on John ja tämä määritellään oikeaksi arvoksi, eli v' = John. Jos arvo v = Jhn, on se silloin väärin ja tästä pystytään laskemaan ero v:n ja v':n välillä. Tätä esimerkkiä kutsutaan syntaktiseksi tarkkuudeksi. Toinen tarkkuuden määrittely on semanttinen tarkkuus. Semanttisessa tarkkuudessa datan pitää kuulua tiettyyn kategoriaa, esimerkiksi henkilönnimi voi olla kategoria. Aiempaa esimerkkiä lainaten, jos v on Michael ja v' on John, on v:n arvo semanttisesti oikein, mutta syntaktisesti väärin. Semanttista tarkkuus määritellään, onko arvo semanttisesti oikein vai ei ja syntaktiselle tarkkuudelle saadaan tarkka matemaattinen arvo. (Batini & Scannapieca 2006, s. 19-22)

Datan täydellisyydellä viitataan siihen, kuinka hyvin data kuvaa täydellistä esitystä siitä tosielämän asiasta, jota ne kuvaavat. Datan täydellisyydessä voidaan siis puhua, puuttuuko saatavilla olevasta datasta tietoa vai ei. Datan täydellisyyttä arvioitaessa on tärkeä tietää, miksi datasta puuttuu tietoa. Tiedon puuttumiselle on useita syitä ja tietoa voi puuttua sen takia, koska tieto on olemassa, mutta sitä ei tiedetä, tai tietoa ei ole olemassa, tai ei tiedetä tarkalleen, onko tietoa olemassa vai ei. Alla oleva taulukko 1. havainnollistaa kuvitteellisen yrityksen asiakastietoja ja siitä voidaan nähdä, että riveillä 2,3 ja 4 sähköposti kentässä on tieto ”NULL”, joka viittaa tiedon puuttumiseen. Tässä esimerkissä henkilön Kalle Korhonen sähköposti saattaa puuttua, koska hänellä ei ole sähköpostiosoitetta ollenkaan, joten tieto ei ole epätäydellistä. Henkilöllä Ville Virtanen on oma sähköpostiosoite, mutta sitä ei ole tiedossa, joten tieto on epätäydellistä. Henkilön Niina Niemisen tapauksessa ei ole tiedossa, onko hänellä omaa sähköpostiosoitetta vai ei, joten ei voida olla varmoja onko tieto epätäydellistä vai ei. Esimerkki havainnollisti myös sitä, että tietoa voi puuttua, mutta se ei tee datasta välttämättä epätäydellistä, pitää vain tietää onko tietoa edes olemassa. (Batini & Scannapieca 2006, s. 23-26)

Taulukko 1. Esimerkki henkilötieto

ID	NIMI	SÄHKÖPOSTI
1	Matti Meikäläinen	Matti.meikalainen@gmail.com
2	Kalle Korhonen	NULL
3	Ville Virtanen	NULL
4	Niina Nieminen	NULL

Datan laadun kolmas dimensio on oikea-aikaisuus ja se voidaan jakaa kolmeen osaan. Oikea-aikaisuuden osat ovat ajantasaisuus, tuoreus ja volatiliteetti. Volatiliteetillä tarkoitetaan, kuinka useasti data muuttuu ajan mukaan. Esimerkiksi henkilön syntymäpäivä ei ole volatiliteettiä, koska se ei muutu ollenkaan, mutta esimerkiksi pörssikurssit ovat hyvin volatiiliä, koska ne muuttuvat hyvin lyhyessä ajassa. Ajantasaisuudella tarkoitetaan sitä, kuinka ajan tasalla käytettävissä oleva data on sen käyttötarkoitusta varten. Ajantasaisessa datassa on otettava huomioon myös se, että se saattaa olla ajan tasalla, mutta se saattaa tulla saataville myöhässä ja tästä syystä ei ole enää käytettävissä sen alkuperäisessä käyttötarkoituksessa. Esimerkiksi, jos yliopiston luentojen ajankohdat ilmoitettaisiin vasta luentojen jälkeen, data olisi ajantasaista, mutta saatavilla myöhässä ja tästä syystä sillä ei ole merkitystä. Datan tuoreudella viitataan

siihen, milloin viimeksi kyseistä dataa on päivitetty viimeksi. Datan oikea-aikasuutta varten on tärkeä varmistaa, että data on ajan tasalla ja se on saatavissa silloin, kun sitä tarvitaan.(Batini & Scannapieca 2006, s. 28-29)
Datan laadun neljäs dimensio on johdonmukaisuus. Johdonmukaisuudella tarkoitetaan sitä, kuinka samanlaisia samaa asiaa kuvaavat datatietueet ovat keskenään. Johdonmukaisuuden rajoitukset asetetaan yleensä tietokantoihin, jotta sinne tallennetut datatietueet olisivat keskenään samanlaisia. Johdonmukaisuuteen vaikuttaa se, kuinka monella eri tavalla sama asia pystytään ilmaisemaan ja tätä pyritään rajoittamaan jo datan keruu vaiheessa. (Batini & Scannapieca 2006, s. 30-33)

Poikkeaviksi havainnoiksi kutsutaan datasta löytyviä havaintoja, jotka eroavat huomattavasti suhteessa muuhun dataan. Esimerkiksi tietokannassa on datajoukko, jonka arvot ovat normaalisti välillä 1-10, mutta yksi arvo on 76. Tästä voidaan nopeasti päätellä, että 76 on poikkeava havainto. Poikkeavia havaintoja syntyy esimerkiksi väärin havainnoidusta, tallennetusta tai väärin tallennetusta datasta tai on kokonaan väärästä paikasta. Poikkeava havainto saattaa myös havainnollistaa harvinaista tapahtumaa. Datan laatua ja poikkeavia havaintoja käsitellessä on tärkeä pyrkiä tunnistamaan, onko datan käsittelyssä tapahtunut virhe vai onko kyseessä harvinainen havainto. Poikkeavat havainnot ovat tapauskohtaisesti tärkeä poistaa, koska ne saattavat aiheuttaa virheitä, kun datasta tehdään analyysejä. Poikkeavia havaintoja pystytään tunnistamaan esimerkiksi laskemalla havainnon etäisyys arvojoukosta, johon sen normaalisti oletettaisiin kuuluvan. (Batini & Scannapieca 2006, s. 86-88)

Poikkeavia havaintoja pystyy myös hyödyntämään, joten ne eivät aina ole haitaksi. Poikkeavien havaintojen löytäminen voi tapauskohtaisesti ilmoittaa esimerkiksi sensorin virheestä, tai poikkeava havainto pankin asiakkaan luottokortin käytössä voi johtua luottokortin päätymisestä vääriin käsiin tai muusta laittomasta toimesta. Poikkeavia havaintoja tunnistavat algoritmit pystyvät pisteyttämään havainnot sen perusteella, kuinka todennäköisesti ne ovat poikkeavia havaintoja tai binäärisesti onko havainto poikkeava vai normaali. (Charu 2015, s. 237-240)

Datan laadussa on tärkeä ottaa myös sen sisältämä kohina. Datan joukossa oleva kohinaksi kutsutaan dataa, joka on tehtävälle analyysille ylimääräistä ja saattaa sisältää poikkeavia havaintoja. Kun dataa kerätään monesta eri lähteestä monessa eri muodossa, on erityisen tärkeää käsitellä ennen sen hyödyntämistä. Datan sisältämä kohina saattaa aiheuttaa datasta

tehtäviin analyyseihin vääriä lopputuloksia. Kohinaa sisältämä data on erittäin haitallista erityisesti koneoppimis algoritmeille, koska se vääristää niiden tarkkuutta ja hidastaa algoritmin toimintaa. Jos kyseessä on reaaliajassa toimiva koneoppimis algoritmi ja käytettävissä oleva data sisältää kohinaa, se aiheuttaa entistä enemmän virheitä staattiseen dataan verrattuna. Kohinaa syntyy dataan ensisijaisesti sensoreista ja ihmisten keräämästä datasta. Kohina pystytään ottamaan huomioon koneoppimis algoritmeissa, jos sen olemassa olosta ollaan tietoisia tai vaihtoehtoisesti dataa voidaan käsitellä, siten että siitä vähennetään kohinan määrää. (Anurag et al. 2018)

3.6 Data-analytiikan hyödyt ja lisäarvo

Tässä kappaleessa käsitellään data-analytiikan potentiaalisia käyttökohteita ja kuinka se voisi tuottaa yritykselle lisäarvoa. Data-analytiikalle löytyy erittäin paljon käyttökohteita, mutta työn rajaamisen puitteissa tässä kappaleessa keskitytään lähtökohtaisesti siihen, kuinka finanssialan toimija pystyy hyödyntämään data-analytiikkaa.

Tähän mennessä on käsitelty datan merkitystä, data-analytiikkaa, big dataa ja koneoppimista. Tässä kappaleessa käydään läpi, kuinka data-analytiikkaa, big dataa ja koneoppimista pystyy hyödyntämään käytännössä ja millaista lisäarvoa yrityksen on mahdollista saada. Yksinkertaisimmillaan data-analytiikalla voidaan vähentää virhealtista manuaalista työtä ja saada tämän pohjalta kustannussäästöjä(TTY Pori 2018, s. 41).

Hyvänä esimerkkinä manuaalisen työn vähentämisestä on JPMorgan Chase pankin vuonna 2016 käyttöönottama järjestelmä, jonka tarkoituksena on lukea ja tutkia lainasopimuksia itsenäisesti. Pankki kertoi uutistoimisto Bloombergille tämän järjestelmän tekevän saman työn murto-osassa ajasta siitä mikä juristeilla ja pankkivirkailijoilla siihen menisi. Tämän järjestelmän avulla JPMorgan Chase säästää vuodessa 360 000 työtuntia. (Kotilainen 2018)
Markkula ja Syväniemi tuovat kirjassaan esille yhden isoimmista data-analytiikan hyödyntäjistä. Heidän mielestään markkinoinnissa hyödynnetään tällä hetkellä analytiikkaa hyvin tehokkaasti. Nykypäivänä analytiikan avulla markkinoinnista on saatu tehtyä personoitua, asiakkaalle kohdistettua, jonka asiakas kokee palveluna. Markkinoinnin suuntana on siirtyä kohti pienempiä kohderyhmiä, reaaliaikaista sisältöä, paikkatiedon ja sosiaalisen median tehokasta hyödyntämistä. Markkula ja Syväniemi (2015, s.120) kertovat

markkinointiautomaation konseptista, jossa data-analytiikka on isossa osassa. Markkinointiautomaatio ei ole uusi konsepti, vaan sen on esitellyt ensimmäisen kerran John D.C. Little vuonna 2001 (Heimbach et al. 2015, s. 130). Little esitti markkinointiautomaation sellaisena prosessina, jossa analysoidaan asiakkaan digitaalista jalanjälkeä, jotta yritykselle saadaan tuotettua merkitsevää informaatiota markkinointi varten (Heimbach et al. 2015, s. 130). Alla olevassa kuvassa 5. on havainnollistettu Markkulan ja Syväniemen esittämä markkinointiautomaation prosessi. Markkinointiautomaation ytimessä on yhdessä toimivat tietojärjestelmät, joka mahdollistaa monikanavaisen kommunikoinnin automatisoinnin ja reaaliaikaisen tulosten mittaamisen. Analytiikan avulla pystytään analysoimaan asiakkaan tuottamaa dataa, joka mahdollistaa asiakkaan tarpeiden tunnistamisen ja ennakoinnin asiakkaan tarpeista. Markkinointiautomaatiolla mahdollistetaan esimerkiksi asiakkaan tietojen keräämisen, segmentoinnin, asiakkaan ostopotentiaalin arvioinnin, kohdistetun ja personoidun viestinnän asiakkaille. Mainostajalle markkinointiautomaatio tekee markkinoinnista kustannustehokkaampaa ja markkinoinnin tehokkuutta on helpompi seurata. (Markkula & Syväniemi 2015, s. 120-127)

Kuva 5. Markkinointiautomaation prosessi (Markkula & Syväniemi 2015, s. 121)

Vuonna 2011 kaksi Hewlet-Packardin (HP) työntekijää Gitali Halder ja Anindya Dey kehittivät analytiikan avulla ennustavan mallin, joka määrittelee jokaiselle 300 000 HP:n työntekijälle todennäköisyyden sille, että he vaihtavat työpaikkaa lyhyen ajan sisään. Tällä tavalla saatiin analysoitua nykyisiä ja uusia työntekijöitä ja vähennettyä työntekijöiden vaihtuvuutta. Vaihtuvuuden väheneminen pienensi henkilökuluja ja teki samalla työyhteisöstä tiiviimmän. Tämän avulla saatiin selville, että työntekijät pysyvät todennäköisemmin yrityksen

palveluksessa, kun työnkuva pysyy mielenkiintoisena, palkka on sopivan suuruinen ja työntekijällä on mahdollisuus saada palkankorotuksia. (Siegel 2016, s. 59-66)

Konsultti yhtiön EY:n ja Forbes Insightin tekemässä tutkimuksessa vuonna 2016 selvisi, että 66 prosenttia yrityksistä, joiden liiketoimintastrategian ytimessä analytiikka on selkeänä osana, ilmoittivat liikevaihdon kasvuksi vähintään 15 prosenttia. Vertauksena yritykset, joilla ei ollut analytiikka liiketoimintastrategiassa, vain 13 prosenttia niistä yrityksistä ilmoittivat liikevaihdon kasvaneen vähintään 15 prosenttia. EY:n ja Forbes Insightin tekemässä tutkimuksessa haastateltiin 1500 yritysjohtajaa suurista yrityksistä, joiden liikevaihto oli vähintään 500 miljoonaa dollaria vuodessa. Tutkimuksen mukaan isoimpana haasteena yrityksissä on ollut saada luotua tarvittava muutos ihmisnäkökulmaan ja organisaation rakenteeseen. Tutkimuksessa selvisi, että edelleen 41 prosentilla yrityksistä on vielä haasteita tehdä yhteistyötä yrityksen IT-osaston, liiketoiminnan ja data-analytiikan osaajien välillä. Yli 70 prosenttia tutkimuksessa menestyneistä yrityksistä hyödyntävät ennakoivaa analytiikkaa ja sisällyttäneet analytiikan osaksi liiketoimintaa. (EY 2017)

Breslow et al. (2017) kertoo artikkelissaan, että koneoppimiseen pohjautuvien tilastollisten mallien avulla voidaan vähentää rahanpesun estämisessä ja transaktioiden valvomisessa virheellisiä-positiivinen hälytyksiä ja manuaalisen työn tarvetta. Mallien avulla voidaan havaita suuresta transaktioiden määrästä tarkemmin epäilyttäviä transaktioita. Breslow et al. (2017) kokemuksesta koneoppimis algoritmien avulla saatiin vähennettyä virheellisten raporttien määrää jopa 20 – 30 prosenttia. Tämän johdosta tutkijat pystyivät keskittymään enemmän tutkivaan työhön ja manuaalisen työn määrä väheni jopa 50 prosenttia. (Breslow et al. 2017)

4 DATA-ANALYTIIKAN HYÖDYNTÄMINEN RAHANPESUN ESTÄMISESSÄ

Tässä kappaleessa käsitellään kirjallisuudesta löydettyjä keinoja data-analytiikan hyödyntämiseen rahanpesun estämisessä. Kappaleen lopussa esitellään myös rahapesun estämiseen tarkoitettuja sovelluksia.

Pankki- ja finanssisektorilla voidaan hyödyntää big data järjestelmiä tunnistamaan luottokortti huijauksia, vakuutus huijauksia ja rahanpesun yrityksiä. Big data järjestelmien avulla dataa

voidaan analysoida useasta lähteestä samanaikaisesti ja siihen voidaan integroida koneoppimis algoritmeja tunnistamaan poikkeavuuksia ja epäilyttäviä transaktioita. Big data analytiikkaa voidaan hyödyntää, myös kun tutkitaan asiakkaan historia tietoja ja erityisesti kun tutkitaan useita transaktioita kerralla. (Bahga & Madisetti 2016, s. 29)

Tällä hetkellä iso osa pankeista hyödyntää sääntöpohjaista järjestelmää tunnistamaan epäilyttäviä transaktioita. Säännöt on määritelty tiettyjen ohjeistuksien mukaan ja tämän pohjalta ei pysty olemaan varma onko kyseessä rahanpesu yritys vai ei. Koneoppimisen hyödyntäminen rahanpesun estämisessä on kustannustehokas vaihtoehto tunnistamaan saatavilla olevasta datasta epäilyttäviä transaktioita. Haasteen koneoppimisen hyödyntämiseen pankki- ja finanssisektorilla on saatavilla olevan datan rajoitteet, joiden avulla koneoppimis algoritmeja voidaan testata. Haasteena on myös ohjelmiston ylläpitämisestä aiheutuvat kustannukset, jotka syntyvät, kun ohjelmistoa päivitetään muuttuneiden vaatimusten pohjalta. Koneoppimis algoritmille haasteita tuottaa puuttuvat tiedot, jotka ovat pankeille yleisiä. Asiakkaasta saattaa puuttua tarpeellisia tietoja, joita asiakas ei ole halunnut yksityisyyden suojan puitteissa luovuttaa. Puuttuva tieto voidaan yrittää korvata erillisten algoritmien avulla tai transaktiot, joihin puuttuva tiedot liittyvät joudutaan poistamaan kokonaan analyysistä. Toimiva rahanpesun tunnistamisjärjestelmä löytää trendejä ja kuvioita transaktioista ja laskee tämän pohjalta todennäköisyydelle, että transaktioissa on kyse rahanpesusta ja tietyn kynnyksen ylittyessä kyseiset transaktiot otetaan tarkempaan tarkasteluun. Data-analytiikan hyödyntämisen tavoitteena on saada vähennettyä virheellisiä-positiivinen hälytyksiä, samalla pitäen huolen, että tapauksien löytämisen tarkkuus pysyy korkeana. Virheellisten-positiivinen hälytyksien vähentäminen vähentää samalla tarvittavia henkilötyötunteja. Koneoppimisen avulla rahanpesun estämisen järjestelmää voitaisiin kehittää siten, että se pystyisi tunnistamaan epäilyttäviä transaktioista täysin uusista tapauksista. Rahanpesun estämisen koneoppimis algoritmit voidaan jakaa kahteen osaan: Ohjattuun- ja ohjaamattomaan oppimiseen. Chen et al. Esittelevät tutkimuksessaan useita eri koneoppimisen algoritmeja, joita voi hyödyntää yksin tai yhdessä toisten tukena rahanpesun estämisen tunnistamisessa. Heidän esittelemiä koneoppimis algoritmeja on esimerkiksi sumea logiikka, k-keskiarvo klusterointi, neuroverkot, päätös puut. Parhaimmat tulokset saatiin yhdistelemällä useita algoritmeja, mutta isoimmaksi haasteeksi ilmeni testattavan datan saatavuus. Suurin osa algoritmeista testattiin vain 10 000 transaktion avulla ja vain muutamaa pystyttiin testaamaan yli miljoonan transaktion avulla. (Chen et al. 2018, s. 246-274)

Palshikar et al. tuovat esille tutkimuksessaan, että rahanpesun estämisessä ja epäilyttävien transaktioiden löytämisessä on käytetty erityisesti ohjaamattoman oppimisen algoritmeja esimerkiksi klusterointia ja anomalioiden havainnointia. Anomalioiksi kutsutaan normaalista poikkeavia kuvioita datassa. Anomalioita voidaan kutsua myös poikkeaviksi havainnoiksi. Monesti klusterointia ja poikkeavien havaintojen analyysiä käytetään yhdessä. Yleinen lähtökohta rahanpesun estämisessä on segmentoida asiakkaiden tilit klustereihin ja etsiä niistä poikkeavia havaintoja hyödyntäen sopivaa yhdenvertaisuus vertailua ja liiketoiminta osaamista. Tämän jälkeen tilit valitaan valvottavaksi asiakas profiilien riskisyyden perusteella ja voidaan verrata, onko valvottavan asiakkaan tili anomalia muihin saman kategorian pankkitileihin verrattuna. (Palshikar et al. 2014, s. 1-4)

Palshikar et al. esittelee tutkimuksessaan heidän kehittelemän rahapesun tunnistamis työkalun, jonka tarkoituksena on yhdistää dataa monesta eri lähteestä. Lähteitä voi olla esimerkiksi pankkien transaktiodataa, vuosikertomuksia ja rahavirtalaskelmia. Tämän jälkeen hyödynnetään analytiikkaa ja työkalu syöttää ulos listan epäilyttävistä toimista ja epäilyttävistä tileistä ja indikaattorit näiden syiksi. (Palshikar et al. 2014, s. 5-6)

Oracle tarjoaa ratkaisua rahanpesun tunnistamiseen tarjoamalla ohjelmistoa, joka ei käytä perinteistä sääntöpohjaista tunnistamista epäilyttävien transaktioiden löytämiseen. Oraclen ohjelmisto käyttää tekoälyä ja koneoppimista tuottamaan riskianalyysiä. Oraclen ohjelmistolla pystyy integroimaan asiakkaan tietoa useasta lähteestä ja analysoimaan näitä lähteitä yhtäaikaisesti ja luomaan niiden pohjalta kattavan riskiprofiilin asiakkaasta. (Oracle 2019)

Oraclen yksi kilpailijoista on SAS. SAS tarjoaa heidän ohjelmistoaan finanssia-alan organisaatioille. SAS ohjelmistossa on Oraclen tapaan tuki monen tietolähteen integroimiselle yhteen paikkaan. SAS ohjelmisto käyttää vertailuryhmien anomalioiden tunnistusta epäilyttävien transaktioiden löytämisessä. Vertailuryhmien anomalioita tutkiessa asiakkaan historiatietoja verrataan nykyiseen käyttäytymiseen ja samalla verrataan toisten samankaltaisten asiakkaiden käyttäytymiseen. SAS ohjelmistolla esimerkiksi pankki pystyy automaattisesti seuraamaan epäilyttävää toimintaa ja samanaikaisesti ohjelmisto dokumentoi omaa päätösprosessiaan ja mahdollistaa esitäytetyt ilmoitukset viranomaisille epäilyttävästä toiminnasta. (SAS 2019b)

5 DATA-ANALYTIIKKA JA ASIAKKAAN TUNNISTAMINEN FINANSSIALAN ORGANISAATIOSSA

Tässä kappaleessa käsitellään tämän työn soveltavaa osuutta. Haastateltavat on esitelty taulukossa 2. Kappaleen lopussa esitetään yhteenveto haastattelun tuloksista.

Taulukko 2. Haastateltavien henkilöiden esitellyt

Henkilö A	Tuoteasiantuntija	2 vuoden kokemus rahanpesun estämisestä ja asiakkaan tuntemisesta. Vastuussa asiakkaan tuntemiseen ja rahanpesun estämiseen liittyvistä prosesseista.
Henkilö B	Head of data science	2.5 vuotta kokemusta head of data science tehtävistä ja vastaa tiimin kanssa tekoälyn ja edistyneemmän analytiikan kehitykse8stä. Yhteensä noin 15 vuoden kokemus tekoälyn ja analytiikan alalta asiantuntijana ja erikokoisten tiimien vetäjänä.
Henkilö C	Anti Financial Crime Specialist	Talousrikollisuuden torjunnan asiantuntija, työssä keskittynyt rahanpesun estämiseen ja työ painottuu tarkemmin analytiikkaan, raportointiin ja järjestelmäkehitykseen.
Henkilö D	Data Scientist	4 vuoden kokemus rahanpesun estämisestä ja data-analytiikasta. Nykyisessä tehtävässä tukee rahanpesutorjunta ja sanktiofunktioita analytiikan avulla.

5.1 Rahanpesun estämisen menetelmät ja haasteet

Henkilön A kuvaa organisaation rahanpesun estämisen menetelmiä vastaavasti: ”*Pankilla on velvollisuus tuntea asiakkaansa ja asiakkaan maksuliikenne. Kun pankki tuntee asiakkaan normaalin maksuliikenteen, on pankin mahdollista erottaa poikkeamat ja epäilyttävät liiketoimet asiakkaan maksuliikenteessä. Asiakkaan maksuliikennettä seurataan sekä*

konttoreissa että keskitetysti." Henkilö A toi esille kolme haastetta rahanpesun estämisessä: " *Asiakkaiden maksuliikenteessä on paljon muutoksia ja aidosti epäilyttävien liiketoimien tunnistaminen on haastavaa", " Ylläpitää pankissa riittävää osaamisen ja tiedon tasoa asiaan liittyen"* ja *"haastavaa on se, miten asiakkaat suhtautuvat pankin tiukentuneisiin velvollisuuksiin".* Henkilö A kertoo, että asiakkaiden valvontaa suoritetaan manuaalisesti erilaisten listojen perusteella. Listoille nousee asiakkaat tiettyjen perusteiden mukaan ja organisaatiossa on tehty päätös ottaa käyttöön rahanpesun estämistä tukeva järjestelmä.

Henkilö B kuvaa rahanpesun estämisen menetelmiä vastaavasti: *"Rahanpesun estämiseen hyödynnetään tyypillisesti asiakkaiden rahaliikenteen ja tilitapahtumien tutkimista sekä asiakkaiden omia ilmoituksia tyypillisestä rahaliikenteestä."* Henkilön B havainnot rahanpesun estämisen haasteita vastaavat hyvin henkilön A:n kokemusta: "*Ongelmana on tunnistaa mikä on tyypillistä tili- ja rahaliikennettä ja mikä on mahdollisesti rahanpesuun liittyvää."* Henkilö B tuo esille sen, että haasteena on eritellä epätyypillinen ja normaali käyttäytyminen, joka mahdollistaisi rahapesu tapauksien selvittämisen ilman ison asiakasmassan seuraamista manuaalisesti. Henkilö B tuo myös esille tiedon saatavuuteen liittyvän haasteen: *"Haasteena on myös yksittäisen pankin kohdalla se, että näkyvyys rahaliikenteeseen rajoittuu kyseisen pankin kautta tapahtuvaan liikenteeseen ja tyypillisesti rahanpesutapauksissa käytetään useita pankkeja tai tilejä."* Henkilön B kertoo myös, että epäilyttävien transaktioiden tunnistamisessa käytetään apuna analytiikkaa- ja riskiennusteita, sekä apuna käytetään asiakkaan itse tekemiä ilmoituksia rahaliikenteestä. Henkilö B kertoo, että tekoälyn ja analytiikan avulla ennustetaan riskitapauksia ja poikkeuksellista rahaliikennettä ja näitä seurataan sen jälkeen manuaalisesti ja epäilyttävät tapaukset lähetetään eteenpäin tutkittavaksi.

Henkilö C kuvaa rahanpesun estämisen pääpiirteet " *Rahanpesun estäminen on pääpiirteiltään transaktioiden monitorointia sekä asiakkaan tuntemistietojen varmistamista."* Henkilö C jakaa transaktioiden monitoroinnin kahteen osa-alueeseen: pakoteseurannan reaaliaikaiseen monitorointiin ja AML-monitorointiin. Henkilö C kuvailee pakoteseurantaa: "*Jokainen lähtevä ja saapuva maksu monitoroidaan järjestelmätuetusti FATF:n asettamia pakotelistoja vastaan*" ja AML-monitoroinnissa: "*Pyritään järjestelmätuetusti löytämään epäilyttäviä toimia tavallisen transaktiomassan seasta."* Henkilö C kertoo, että asiakkaan tunteminen-osasto toimii yhteistyössä AML-monitoroinnin ja asiakkaan tunteminen-osaston tehtävänä on: "*varmistaa, että tunnemme asiakkaidemme ja kirjeenvaihtajapankkimme toiminnan ja*

vastuuhenkilöt tarpeeksi hyvin, jotta asiakkainamme ei ole sellaisia osapuolia, joita emme riskinottohalukkuuden puitteissa haluaisi säilyttää asiakkaina." Henkilö C tuo rahanpesun estämisen haasteiksi ensimmäisenä esille haasteen löytää isosta transaktiomassasta ne epäilyttävät transaktiot, haasteena on myös määrittää epäilyttävät transaktiot. Henkilö C kertoo viranomaisten asettamista vaatimuksista: "*Viranomaisten asettamat vaatimukset eivät määrittele millaisia raja-arvoja tai menetelmiä pankkien tulisi käyttää ja mitkä transaktiot tulisi ilmoittaa rahanpesun selvittelykeskuskeskukselle. Käytännössä ohjeistus on "Ilmoittakaa kaikki epäilyttävä" ja pankkien tulee ratkaista, mikä on tarpeeksi epäilyttävää heidän kyvykkyyksien, järjestelmien ja riskinottohalukkuuden puitteissa."* Henkilö C tuo myös esille sen, että rahanpesun selvittelykeskus haluaisi mahdollisimman vähän aiheettomia ilmoituksia, mutta ei kerro ilmoittajalle mitkä ilmoitukset ovat johtaneet rahanpesu epäilyn tutkintaa. Henkilön C organisaatiossa asiakkaiden toimintaa valvotaan useilla eri järjestelmillä automatisoidusti, mutta prosessiin kuuluu edelleen paljon manuaalista työtä, kuten rahojen alkuperän selvittämistä ja asiakkaan tuntemistietojen päivittämistä.

Henkilö D kokemusta tärkein asia rahanpesun estämisessä on asiakkaan tunteminen ja tilitapahtumia yleisesti peilataan asiakkaan profiiliin ja tämän jälkeen katsotaan, onko informaatio poikkeavaa. Henkilö D kokemuksesta suurimmat haasteet rahanpesun estämisessä on: "*Datan laatu, saatavuus ja eri tietolähteiden yhdistäminen on suuri haaste. Periaatteessa pystyt rakentamaan mitä tahansa profiileja ja pyöritellä kaikennäköistä analytiikkaa, mutta se underlying data asettaa rajoitukset tähän."* Henkilön D organisaatiossa on käytössä maksuliikenteen monitorointijärjestelmä, jonka avulla pyritään tunnistamaan epäilyttävät tilitapahtumat. Henkilö D tuo esille sen, että epäilyttävien tilitapahtumien tarkistaminen on manuaalista työtä ja vanhantyyppiset monitorointi järjestelmät tuottavat paljon virheellisiä-positiivinen hälytyksiä. Henkilö D ehdottaa, että virheellisiä-positiivinen hälytyksiä voisi vähentää ohjelmistorobotiikalla tai koneoppimisen avulla.

5.2 Asiakkaan tunnistaminen

Henkilö A kertoo, että asiakkaat tunnistetaan virallisen henkilötodistuksen kanssa tai asiakkaat voivat tunnistautua myös sähköisesti verkkopankkitunnuksilla. Henkilö B kertoo, että asiakkaat tunnistetaan hyvin samalla tavalla kuin Henkilö A. Henkilö B kertoo asiakkaan tunnistamisosaston työnkuvasta: *"Heidän on arvioitava asiakkaan riskisyys ja tehtävä*

jatkuvasti tarkistuksia ovatko tiedot tai riskitaso muuttuneet. Jos asiakas havaitaan korkean riskin asiakkaaksi, edellyttää tämä asiakkaan tehostettua tuntemista"

5.3 Rahanpesun estämisen aiheuttamat kulut

Kaikkien vastanneiden kesken yhteisenä kuluna rahanpesun estämisessä oli henkilötyötunneista koituvat kustannukset. Epäilyttävän liiketoiminnan tarkastus ja ilmoitusten tekeminen vaatii edelleen manuaalista työtä ja se on iso kuluerä. Henkilö A toi esille myös tulevat ohjelmistoinvestoinnit tulevat lisäämään kuluja. Henkilö B organisaatiossa kuluja syntyy rahanpesun estämisen järjestelmien kehittämisestä ja ylläpidosta Henkilö C kertoo, että rahanpesun estämisen kulut liikkuvat nykypäivänä miljoonissa ja suuri osa kuluista syntyy järjestelmien lisensseistä ja kehittämisestä.

5.4 Asiakastietoihin liittyvät rajoitukset

Henkilön A:n mielestä uuden tietosuoja-asetuksen johdosta pankin tulee arvioida entistä tarkemmin peruste jokaiselle asiakkaalta kysytylle tiedolle. Samalla tietosuoja-asetus on rajoittanut asiakkaasta tallennettavaa tietoa, koska pankki voi tallentaa vain sellaista tietoa, johon pankilla on painava peruste. Henkilön A kertoo, että asiakkaalta ei voida ihan kaikkea kysyä, vaikka se voisi olla eduksi rahanpesun estämisessä. Henkilön B mielestä tietosuoja-asetus ei ole hankaloittanut asiakkaan tunnistamis prosessia, koska: *"Rahanpesulainsäädäntö on viranomaismääräys, joka mahdollistaa asiakkaan tietojen käsittelyn tähän käyttötarkoitukseen. "* Henkilön C mielestä tietosuoja-asetus ei ole hankaloittanut asiakkaan tunnistamista, mutta on aiheuttanut kustannuksia järjestelmäkehitykseen ja tietovarastointiin. Henkilö C kertoo myös, että *"Teoriassa asiakkaista on vähemmän tietoa saatavilla eri lähteistä, sillä yksityishenkilöt pyytävät yrityksiä poistamaan heidän tietonsa nopeammin kuin aiemmin ja myös tietojen poistorutiinien tahti on nopeutunut."* Henkilö C kertoo, että asiakastietoihin liittyy rajoituksia tiedon jatkuvuuteen liittyen. Tietokantoja on puhdistettava ajoittain, joka hankaloittaa koko asiakashistorian analysoimista.

5.5 Data-analytiikka finanssialan organisaatiossa

Henkilö A kertoo, että heillä ei tällä hetkellä ole data-analytiikkaa hyödyntävää järjestelmää, mutta ovat sellaisen hankkimassa. Henkilö A:n mielikuva data-analytiikan hyödyntämisestä on,

että *”Data-analytiikan avulla organisaatio pystyy käsittelemään isoa määrää dataa tehokkaasti. Pankilla on paljon tietoa asiakkaasta ja asiakkaan maksuliikenteestä, joten on tärkeää, että järjestelmä ”pureskelee” tiedosta olennaisimman esiin. Tällöin säästetään henkilöresursseja ja pystytään tarkkailemaan tehokkaasti asiakkaita ja asiakkaiden maksuliikennettä lain vaatimalla tavalla.”*

Henkilö B:n organisaatiossa *”Data-analytiikkaa hyödynnetään kaikkiin pankin prosesseihin asiakaskokemuksen ja parempien ja helpompien palveluiden luomiseksi.”* Henkilön B mielestä data-analytiikalla voidaan *”Parantaa ymmärrystä prosesseista ja asiakkaista, automatisoida manuaalisia vaiheita ja lisätä prosessien tehokkuutta ja nopeutta.”*

Henkilön C:n mielestä data-analytiikasta on erityisesti rahanpesun estämiseen paljon hyötyä ja hän kertoo, kuinka data-analytiikkaa hyödynnetään: ”*Suuren datamäärän ansiosta pystymme esimerkiksi laskemaan, kuinka todennäköisesti tietynlaisen henkilön toimi tulee johtamaan rahanpesuilmoitukseen. Pystymme myös koneoppimismalleja hyödyntämällä ehkä tulevaisuudessa sulkemaan osan järjestelmien luomista hälytyksistä, ilman, että se vaatisi tutkintatyötä. Analytiikalla voidaan mm. resursoida henkilöstön oikein, kehittää uusia transaktiomonitoroinnin skenaarioita, estimoida kuinka moni epäilyttävä transaktio jää tutkinnan ulkopuolelle sekä opettamaan algoritmeja kehittämään uusia rahanpesuskenaarioita tai sulkemaan rahanpesuhälytyksiä.”* Henkilön C:n mielestä data-analytiikan tulisi olla kaikille yrityksille, ketkä ovat ilmoitusvelvollisia itsestään selvyys. Henkilö C toi kuitenkin esille data-analytiikan käyttämisen haasteita: ”*Näen kuitenkin haasteena sen, että pankilla on käytettävissä vain oma data. Kaikki pankit toimivat talousrikollisuuden alueella kuitenkin lähes samalla tavalla ja kyseinen osa-alue ei ole kilpailullinen, joten näen siinä paljon potentiaalia, jos pankit ja viranomaiset voisivat jakaa dataa keskenään ja toimia tiiviimmin yhteistyössä. Dataa voisi jakaa pankkien kesken esimerkiksi lohkoketju-teknologiaa hyödyntämällä, jolloin varmistettaisiin, että oikeilla osapuolilla on pääsy dataan. Jos pankit saisivat tietoonsa, mitkä rahanpesuilmoitukset ovat johtaneet tutkintaan ja/tai oikeuden päätökseen, voisimme opettaa algoritmia entistä paremmin poimimaan tämän tyyppiset asiakkaat jo siinä vaiheessa, kun heidät ollaan ottamassa asiakkaaksi.”*

Henkilön D mielestä *”pankilla on lähtökohtaisesti paljon dataa ja toimiala on hyvin datakeskeinen. Datan tärkeys ja sen hyödyntäminen varsinkin bisneksen teossa ja arvonluonnissa tulee varmasti kasvamaan tulevaisuudessa.”* Henkilön D mielestä on tärkeä määritellä mitä halutaan tehdä ja mitä voidaan, jotta data-analytiikalla voidaan saada lisäarvoa.

”Kaiken tyyppisiä lukuja ja statistiikkoja voidaan aina tuottaa ja pyöritellä, mutta jos business value on nice to know -tasolla, niin perimmäistä kysymystä on hyvä iteroida pari kertaa.”

5.6 Haastattelujen yhteenveto

Tässä kappaleessa esitetään haastattelujen yhteenveto jokaisen kysymyksen kohdalta taulukossa 3.

Taulukko 3. Haastattelujen yhteenveto

Rahanpesun estämisen menetelmät ja haasteet	Jokaisen vastaajan vastaukset olivat hyvin samanlaiset ja jokainen vastaaja toi esille heidän velvollisuutensa tunnistaa ja tuntea heidän asiakkaansa toiminta ja maksuliikenne. Yhdistävänä tekijänä oli myös, että asiakkaan tilitapahtumia käytetään lähtökohtana rahanpesun tunnistamisessa. Haasteiksi kolme vastaajaa neljästä kertoivat asiakkaan maksuliikenteen analysoimisen ja sieltä löytää epäilyttävää toimintaa. Haasteeksi ilmeni myös datan laatu, saatavuus ja tietolähteiden integroiminen. Isoksi haasteeksi ilmeni se, että rahanpesu toiminnan kehittäminen tarkemmaksi on haastavaa datan rajallisuuden takia.
Asiakkaan tunnistaminen	Asiakkaan tunnistamisessa oli kaikilla vastaajilla hyvin samanlaiset näkökulmat, eli asiakas otetaan vastaan ja tunnistetaan henkilöllisyys ja tämän jälkeen luodaan asiakkaasta riskiprofiili. Asiakkaan valvomisessa oli eroja haastateltavien välillä. Yhdistävänä tekijänä oli se, että kaikkien haastateltavien organisaatiossa työ vaatii edelleen manuaalista työtä, mutta haastateltavien kesken työkalujen ja automaation käyttämisessä oli huomattavia eroa.
Rahanpesun estämisen aiheuttamat kulut	Vastaajien kesken oli tähän kysymykseen hyvin samanlaiset vastaukset. Kuluja syntyy lähtökohtaisesti henkilöstö ja järjestelmien kehittämisestä.
Asiakastietoihin liittyvät rajoitukset	Tähän kysymykseen vastaajilla oli vaihtelevia vastauksia. Henkilön A:n mielestä EU:n yleinen tietosuoja-asetus on

	vaikuttanut asiakkaan tuntemisprosessiin, mutta Henkilön C:n mielestä tietosuoja-asetus ei ole merkittävästi hankaloittanut tuntemisprosessia. Kaikkien vastanneiden kesken yhtenäistä oli, että tietosuoja-asetus on rajoittanut saatavilla olevaa tietoa.
Data-analytiikka finanssialan organisaatiossa	Kolmella vastanneella oli heidän organisaatiossansa käytössä data-analytiikkaa hyödyntävä ratkaisu tai ratkaisuja rahanpesun estämisessä. Yhdellä vastanneesta data-analytiikka ratkaisu oli vasta hankinnassa. Kaikilla vastanneista oli hyvin samanlainen mielikuva siitä, mitä hyötyä finanssialan organisaatio voisi saada hyödyntämällä data-analytiikkaa.

6 JOHTOPÄÄTÖKSET

Tämän kandityön tarkoituksena oli tutkia, kuinka data-analytiikkaa voitaisiin käyttää hyväksi finanssialan organisaatiossa ja erityisesti, kuinka data-analytiikkaa voitaisiin käyttää apuna rahanpesun estämisessä. Tutkimuksen alussa syvennyttiin, kuinka lainsäädäntö ohjaa finanssialan organisaatioita rahanpesun estämisessä, sekä millaisia haasteita ja riskejä rahanpesun estäminen aiheuttaa finanssialan organisaatiolle. Tutkimuksessa siirryttiin seuraavaksi data-analytiikkaan ja syvennyttiin, dataan ja sen laatuun, big dataan ja koneoppimiseen. Tutkimuksessa käytettiin teoriakirjallisuutta ja artikkeleita apuna selvittämään, kuinka tällä hetkellä data-analytiikkaa hyödynnetään rahanpesun estämisessä, sekä tutkittiin mitä haasteita teoriakirjallisuudessa oli mainittu data-analytiikan hyödyntämiseen. Teoriakirjallisuuden on samalla tarkoitus antaa lukijalle tarvittava tietämys data-analytiikasta ja rahanpesun estämisestä. Haastattelujen perusteella tunnistettiin asiakkaan tunnistamiseen, asiakkaan tuntemiseen ja rahapesu epäilyksien löytämisen haasteita. Haastattelusta saadun tiedon ja teoriakirjallisuudesta saadun tiedon avulla data-analytiikalle löydettiin käyttökohteita.

Tutkimuksen ensimmäinen tutkimuskysymys oli:

1. Millaista lisäarvoa data-analytiikan hyödyntäminen voi tuoda finanssialan organisaatiolle?

Teoriakirjallisuudessa tuli esille, että data-analytiikalle on finanssialan organisaatiolle lukuisia käyttökohteita, eikä ne rajoitu pelkästään rahanpesun estämiseen. Finanssialan organisaatio pystyy hyödyntämään data-analytiikkaa hyvin myös omissa sisäisissä prosesseissa ja vähentämään työntekijöiltä vaadittavaa manuaalista työtä. Erityisesti kirjallisuudessa korostettiin, että data-analytiikan avulla voidaan myös vähentää ihmisten virheistä aiheutuvia kustannuksia. Data-analytiikan ei tarvitse korvata ihmistyövoimaa, vaan se voisi olla tärkeänä työkaluna tehostamassa työntekijän toimenkuvaa. Data-analytiikan avulla pystytään käsittelemään automaattisesti moninkertaisesti enemmän dataa ja jalostamaan siitä informaatiota nopeammin, kuin normaali työntekijä. Data-analytiikan avulla finanssialan organisaatiot pystyvät esimerkiksi tehostamaan omaa markkinointia automatisoimalla sen. Data-analytiikka työkalujen käyttöönotossa on hyvä kartoittaa saatavilla oleva data ja varmistaa sen laatu, koska analyysin lopputulos on yhtä hyvä, kuin analyysissä käytetyn datan laatu.

Tutkimuksen toinen tutkimuskysymys oli:

2. Miten data-analytiikkaa voidaan hyödyntää rahanpesun estämisessä?

Teoriakirjallisuudesta löytyi erityisesti koneoppimiseen liittyviä algoritmeja, joita voitaisiin hyödyntää rahanpesun estämisessä ja erityisesti epäilyttävien transaktioiden tunnistamisessa. Isona haasteena näissä algoritmeissa oli testattavan datan puute. Koneoppimisalgoritmien toimivuuden testaamista varten tarvitaan useasti hyvinkin paljon dataa, jotta voidaan varmistaa sen skaalautuvuus, toimivuus ja tarkkuus. Kirjallisuudessa tuli ilmi, että oikean transaktiodatan ja oikeiden rahanpesu tapaus tietojen saaminen koitui haastavasti. Haastatteluissa tuli ilmi, että pankeille ei ilmoiteta jälkikäteen, onko heidän ilmoituksensa pohjalta aloitettu rahanpesu epäilyn tutkinta. Toinen iso datan puuttuvuuden ongelma on, että pankilla on käytössään pelkästään omaa dataa. Rahanpesun estämisen kannalta ja algoritmien kehittämisen kannalta olisi kaikille osapuolille parasta, jos pankit voisivat käyttää yhdessä asiakkaiden dataa. Erityisesti sellaisessa tilanteessa, että asiakkaalla on asiakassuhde usean pankin kanssa. Samalla finanssivalvonta pystyisi tekemään enemmän yhteistyötä pankkien kanssa ja ilmoittamaan

tarkemmin, mitkä ilmoitukset ovat johtaneet tutkintaan ja tämän perusteella koneoppimis algoritmeja pystyttäisiin kehittämään tarkemmiksi. Tutkimuksessa käsiteltiin myös valmiita ohjelmistoja, jotka hyödyntävät data-analytiikkaa rahanpesun estämisessä. Nämä ohjelmistot hyödynsivät samoja tekniikoita, mitä teoriakirjallisuudessa tuli esille. Haastatteluissa tuli esille, että useassa organisaatiossa hyödynnetään jo data-analytiikkaa ja data-analytiikka ohjelmistoja rahanpesun estämisessä.

Jatkokehityksen kannalta mielestäni olisi tärkeää selvittää viranomaisten kanssa olisiko mahdollista toteuttaa rahanpesun esimerkkitapauksia sisältävä tietokanta, jota pankit ja kehittäjät pystyisivät hyödyntämään suunnitellessa ja testatessaan uusia rahanpesun vastaisia järjestelmiä. Tietokanta voisi sisältää oikeaa dataa epäilyttävistä transaktioista ja rahanpesu tapauksista, mutta data olisi sellaisessa muodossa, että sen alkuperäistä lähdettä tai omistajaa ei voisi selvittää. Tietokanta voisi sisältää myös normaaleja transaktioita ja dataa, joka ei sisällä epäilyttäviä transaktioita, eikä rahanpesu epäily tapauksia. Tietokanta voisi olla esimerkiksi finanssivalvonnan ylläpitämä ja he voisivat tarjota tietokantaan rajapinnan, jonka kautta pankit ja kehittäjät pystyvät sitä käyttämään.

LÄHTEET

Ahonen T. & Kortelainen H. & Kunttu S. (2017). Teollinen internet uudistaa palveluliiketoimintaa ja kunnossapitoa. Teoksessa: Martinsuo, M., & Kärri, T. (Toimittajat) Kunnossapitoyhdistys ProMaint. s. 15 - 25 ISBN 978-952-68687-0-7.

Ahsan, S & Shah, A. Data, Information, Knowledge, Wisdom: A Doubly Linked Chain. Research and Development Center of Computer Science University of Engineering and Technology, Lahore [WWW-dokumentti] [Viitattu 14.10.2019] Saatavissa: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.89.5378&rep=rep1&type=pdf

Anurag & Rani, S. & Rao, S. 2018. Study and Analysis of Noise Effect on Big Data Analytics. International Journal of Management, Technology and Engineering. Vol. 8:7 s. 5841 – 5850. ISSN 2249-7455

Bahga, A. & Madisetti, V. 2016. Big Data Analytics: A Hands-On Approach. Arshdeep Bahga & Vijay Madisetti.

Batini, C. & Scannapieca, M. 2006. Data Quality Concepts, Methodologies and Techniques. Springer Berlin Heidelberg New York

Bonnin, R. 2017. Machine Learning for Developers. Birmingham: Packt Publishing.

Breslow, S. & Hagstroem, M & Mikkelsen, D & Robu, K. 2017. The new frontier in anti–money laundering. [WWW-dokumentti] [Viitattu 3.12.2019] Saatavissa: https://www.mckinsey.com/business-functions/risk/our-insights/the-new-frontier-in-anti-money-laundering

Brännare, S. 2018. Danske Bankin osake romahtanut – Edessä voivat olla miljardikorvaukset, kun Yhdysvallat aloitti rahanpesun tutkinnan. [WWW-dokumentti] [Viitattu 2.12.2019] Saatavissa: https://yle.fi/uutiset/3-10439003

Buyya, R., Calheiros, R. N. & Dastjerdi, A. 2016. Big data: principles and paradigms. Cambridge, MA: Elsevier/Morgan Kaufmann.

Charu C. 2015. Data mining: the textbook. New York, NY: Springer Science Business Media. ISBN 978-3-319-14141-1

Dataquest. 2019. Data Science Terms and Jargon: A Glossary. [WWW-dokumentti] [Viitattu 10.12.2019] Saatavissa https://www.dataquest.io/blog/data-science-glossary/

Davenport, T. H. & Harris, J. G. 2007. Analysoi ja voita. Kilpailun uusi tiede. Helsinki: Talentum.

Devi, V.S & Murty, M.N. 2015. Introduction to Pattern Recognition and Machine Learning. World Scientific Publishing Co. Pte. Ltd. Singapore.

Earley, C. E. 2015. Data analytics in auditing: Opportunities and challenges. *Business Horizons*, 58(5), pp. 493-500.

Exasol. 2019. Definition of Big Data. [WWW-dokumentti] [Viitattu 16.12.2019] Saatavissa: https://www.exasol.com/en/insights/big-data-and-analytics-glossary/big-data/

Finanssivalvonta. Tietoa Finanssivalvonnasta. [WWW-dokumentti] [Viitattu 16.12.2019] Saatavissa: https://www.finanssivalvonta.fi/finanssivalvonta/

Finanssivalvonta. 2019. Finanssivalvonta on määrännyt seuraamusmaksun S-Pankki Oy:lle sekä antanut julkisen varoituksen FIM Varainhoito Oy:lle laiminlyönneistä asiakkaan tuntemisessa. [WWW-dokumentti] [Viitattu 16.12.2019] Saatavissa: https://www.finanssivalvonta.fi/tiedotteet-ja-julkaisut/lehdistotiedotteet/2019/finanssivalvonta-on-maarannyt-seuraamusmaksun-s-pankki-oylle-seka-antanut-julkisen-varoituksen-fim-varainhoito-oylle-laiminlyonneista-asiakkaan-tuntemisessa/

Finanssivalvonta. 2015. Asiakkaan tunteminen – rahanpesun ja terrorismin rahoittamisen estäminen. [WWW-dokumentti] [Viitattu 1.12.2019] Saatavissa: https://www.finanssivalvonta.fi/globalassets/fi/saantely/maarayskokoelma/standardit/2.4/2.4.std6.pdf

Finlex. 2017. Laki rahanpesun ja terrorismin rahoittamisen estämisestä. [WWW-dokumentti]. [Viitattu 1.12.2019]. Saatavissa: https://www.finlex.fi/fi/laki/ajantasa/2017/20170444#L1

Frankenfield, J. 2019 Data-analytics. [WWW-dokumentti] [Viitattu 10.12.2019] Saatavissa: https://www.investopedia.com/terms/d/data-analytics.asp

Gupta, Bhasker. 2016. Interview questions in business analytics. Apress, Berkley, CA.

Heimbach, I. & Kostyra, D.S. & Hinz, O. (2015). Marketing Automation. Business & Information Systems Engineering, Vol. 57:2, s.129–133.

HE 25/2008 Hallituksen esitys Eduskunnalle laiksi rahanpesun ja terrorismin rahoittamisen estämisestä ja selvittämisestä sekä eräiksi siihen liittyviksi laeiksi [WWW-dokumentti] [Viitattu 1.8.2019] Saatavissa: https://www.finlex.fi/fi/esitykset/he/2008/20080025?search%5Btype%5D=pika&search%5Bpika%5D=HE%2025%2F20082

Hirsjärvi, S. & Hurme, H. 2015. Tutkimushaastattelu: Teemahaastattelun teoria ja käytäntö. Helsinki: Gaudeamus Helsinki University Press.

Ikävalko, K. 2019. Suomessa otetaan käyttöön uusi "häpeäpaalu" rahanpesun ehkäisyssä – Rikollinen raha voi jäädä pimentoon jopa tavallisissa autokaupoissa. [WWW-dokumentti] [Viitattu 1.12.2019]. Saatavissa: https://yle.fi/uutiset/3-11024159

Kotilainen, S. 2018. Tekoälyn vallankumous on alkanut – tätä kaikkea se tarkoittaa. [WWW-dokumentti] [Viitattu 28.11.2019] Saatavissa: https://www.tivi.fi/uutiset/tekoalyn-vallankumous-on-alkanut-tata-kaikkea-se-tarkoittaa/f430ff4c-5427-30bd-bdff-8df678315521

Louridas, P. & Ebert, C. 2016. Machine Learning. IEEE Software, Vol. 33:5, s.110–115.

Markkula, T. & Syväniemi, A. 2015. Analytiikkamatka: Datasta tietoon ja tiedolla johtamiseen. Helsinki: Suomen Liikekirjat.

Marr, Bernard. 2018. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. [WWW-dokumentti] [Viitattu 1.8.2019] Saatavissa: https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#3be95a0e60ba

Oracle. Staying Ahead of Financial Crime. [WWW-dokumentti] [Viitattu 1.12.2019] Saatavissa: https://www.oracle.com/industries/financial-services/analytics/network-analytics-graph-intelligence-aml-compliance.html

Palshikar, G.K. & Apte, M. & Baskaran, S. (2014). Analytics for Detection of Money Laundering s.10

Rastas, T. & Asp, E. 2014. Big datan hyödyntäminen. Liikenne- ja viestintäministeriön julkaisuja 20/2014.

Rahanpesulaki 1-8 luku. Laki rahanpesun ja terrorismin rahoittamisen estämisestä. (28.6.2017/444)

Reese, B. 2018. GDPR and EU AML Directives – A Regulatory Tug-of-War? [WWW-dokumentti] [Viitattu 2.12.2019] Saatavissa: https://blog.protiviti.com/2018/05/24/gdpr-eu-aml-directives-regulatory-tug-war/

Rikoslaki 32 luku 6-10 § Kätkemis- ja rahanpesurikoksista (31.1.2003/61)

Rouse, M. 2016. Data analytics (DA). [WWW-dokumentti] [Viitattu 14.10.2019] Saatavissa: http://searchdatamanagement.techtarget.com/definition/data-analytics

SAS. 2019a. Analytics – What it is and why it matters. [WWW-dokumentti] [Viitattu 19.12.2019] Saatavissa: https://www.sas.com/en_us/insights/analytics/what-is-analytics.html#industries

SAS. 2019b. SAS ANTI-MONEY LAUNDERING - Monitor suspicious activity. Make fast decisions. And stay in compliance. [WWW-dokumentti] [Viitattu 1.12.2019] Saatavissa: https://www.sas.com/en_us/software/anti-money-laundering.html

Siegel, E. 2016. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. New Jersey; Hoboken: John Wiley & Sons, Inc.

Sullivan, K. 2015. Anti–Money Laundering in a Nutshell: Awareness and Compliance for Financial Personnel and Business. Berkeley, CA: Apress.

Tarvainen, Ida-Ellen. 2019. Rahanpesuindikaattorit. Keskusrikospoliisi. Rahanpesun selvittelykeskus. [WWW-dokumentti] [Viitattu 1.12.2019] Saatavissa: https://www.poliisi.fi/instancedata/prime_product_julkaisu/intermin/embeds/poliisiwwwstructure/77455_Rahanpesuindikaattorit_11.1.2019_FINAL.pdf?e01ed3f69dfbd688

TTY Pori / Analyyttinen-hanke. 2018. Selvitys data-analytiikan nykytilasta ja data-analytiikan hyödyntämisestä satakunnassa. [WWW-dokumentti] [Viitattu 28.11.2019]Saatavissa: http://www.datatiede.fi/wp-content/uploads/2018/10/Data-analytiikan-selvitys-Julkaisuversio-2-2018-10-30.pdf

UNODC, Illicit money how much is out there, 2011 [WWW-dokumentti] [Viitattu 10.10.2019] Saatavissa: https://www.unodc.org/unodc/en/frontpage/2011/October/illicit-money_-how-much-is-out-there.html

Valtiovarainministeriö, Rahanpesun estäminen rahoitusmarkkinoilla, 2019 [WWW-dokumentti] [Viitattu 10.10.2019] Saatavissa: https://vm.fi/rahanpesun-estaminen-rahoitusmarkkinoilla

Watson, H. J. 2012. The Necessary Skills for Advanced Analytics, Business Intelligence Journal. Vol. 17:4, s. 4-7.

Xanthopoulos, P. & Pardalos, P. & Trafalis, T. (2013). Robust data mining. New York; London: Springer. ISSN 2191-575X (sähköinen)

Chen, Z., Van Khoa, L.D., Teoh, E.N., Nazir, A., Karuppiah, E.K. and Lam, K.S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowledge and Information Systems*, Vol. 57:2, s.245–285.

LIITTEET

Liite 1: Haastattelurunko

Kerrotko itsestäsi ja työtehtävistäsi?

Mitä menetelmiä rahanpesuun estämiseen liittyy?

Mitä haasteita rahanpesun estämiseen tuo?

Millaisilla menetelmillä tunnistatte asiakkaat?

Millaisilla menetelmillä asiakkaita valvotaan ja onko näitä prosesseja automatisoitu, vai liittyykö niihin manuaalista työtä?

Millaisia kuluja syntyy rahanpesun estämisestä?

Millaisia rajoituksia on asiakkaasta saataviin tietoihin?

Onko esimerkiksi EU:n yleinen tietosuoja-asetus (GDPR) hankaloittanut asiakkaan tunnistamisprosessia?

Onko asiakkaasta saatavissa tiedoissa käyttörajoituksia?

Millaista kokemusta sinulla on data-analytiikasta? Onko teillä hyödynnetty data-analytiikkaa yrityksessä aiemmin, jos on niin miten?

Millaista lisäarvoa data-analytiikka voisi mielestäsi tuoda yritykselle