Sami Salminen

BIG DATA -ILMIÖ, SUURET DATAMASSAT

Tietojenkäsittelyn koulutusohjelma

2019

samk
Satakunnan ammattikorkeakoulu
Satakunta University of Applied Sciences

BIG DATA -ILMIÖ, SUURET DATAMASSAT

Salminen, Sami
Satakunnan ammattikorkeakoulu
Tietojenkäsittelyn koulutusohjelma
Lokakuu 2019
Sivumäärä: 41
Liitteitä: 1

Asiasanat: Big Data, pilvipalvelut, 3V malli

Tämän opinnäytetyön tarkoituksena oli tehdä lyhyt yleisesitys niin sanotusta big datasta. Työssä tarkasteltiin big datan määritelmää sekä 3V-mallia. Big datan louhinnassa kerrottiin, mitä louhimisella tarkoitetaan ja miten koneoppimisen metodeja ja algoritmejä hyödynnetään tiedon saamiseksi. Datatyypeistä selvitettiin rakenteettoman, rakenteellisen ja puolirakenteellisen datan ominaisuuksia ja suhteita toisiinsa sekä datan analysointimalleja. Opinnäytetyössä käsiteltiin lisäksi, miten datamassoista saadusta informaatiosta tulee tietämystä.

Big dataa hyödyntävistä alueista julkishallintoa, liike-elämää ja terveydenhuoltoa käsiteltiin lisäksi esimerkein. IoT-laitteista ja roboteista esiteltiin valikoituja käyttökohteita.

Opinnäytetyössä käsiteltiin luonnollisesti myös pilvipalveluiden määritelmää, joitakin pilvipalveluntarjoajien ratkaisuja big datan käsittelyyn ja lyhyesti pilvipalveluiden tietoturvaa sekä sitä, miten EU:n yleinen tietosuoja-asetus vaikuttaa kuluttajiin ja organisaatioihin.

Opinnäytetyön tuloksena saatiin yleiskuva big datasta, miten dataa syntyy, big datan määritelmästä, louhinnasta, datatyypeistä, miten datasta tulee tietämystä sekä joistakin käyttökohteista ja pilvipalveluista.

The Big Data Phenomenon and Large Data Bases

Salminen, Sami
Satakunnan ammattikorkeakoulu, Satakunta University of Applied Sciences
Degree Programme in Business Information Technology
October 2019
Number of pages: 41
Appendices: 1

Keywords: Big Data, cloud services, Three V Model

The purpose of this thesis was to give its readers a brief review of so-called Big Data. The thesis covered, inter alia, the definition of Big Data and the important Three V Model. When discussing data-mining, this term was first defined, and thereafter the readers were told how the methods and algorithms of machine learning can be used to obtain useful information. Moreover, core properties of the three data categories of unstructured, structured and semi-structured data as well as their internal relationships were discussed. The thesis also dealt with the issues of deriving knowledge out of the information obtained from the mining processes.

The thesis presented public services, business enterprises and healthcare systems as examples of practical areas that have already shown to profit highly in various ways from the use of Big Data. Moreover, select areas of use of IoT appliances and robots were briefly ventilated in the thesis.

As a natural necessity the thesis also dealt with the definition of cloud services as well as various solutions given by a number of cloud service providers to handle Big Data, and – not unexpectedly – security issues inherent to cloud services, where the thesis also took up the role of the European Union data protection regulation (General Data Protection Regulation) and its effects on citizens and organizations.

The thesis produced as its major outcome an overall view of Big Data as well as a view on the uprising of data, the definition of Big Data, its mining and its data types, and how data can be developed into knowledge. Finally some areas of use of Big Data as well as cloud services were ventilated.

.

SISÄLLYS

1 JOHDANTO

Big datan suuren tietomassan jatkuva ja nopea lisääntyminen aiheuttaa haasteita datan tallentamiseen ja käsittelyyn. Tämän opinnäytetyön tarkoituksena on tehdä pieni tietopaketti big datasta ja sen käytöstä.

Tässä työssä kerrotaan lyhyesti big dataan liittyvistä osa-alueista: Mitä big data on ja joistakin sen ominaisuuksista, louhinnasta, datatyyppien rakenteista, analysointimalleista ja miten kerätystä datasta saadaan tietämystä. Big dataa hyödyntävät alueet rajataan tässä työssä aiheen laajuuden vuoksi julkishallintoon, liike-elämään, terveydenhuoltoon ja IoT-laitteita käyttäviin kulkuneuvoihin, älykoteihin ja älykaupunkeihin sekä siihen miten robotteja hyödynnetään autoteollisuudessa, maataloudessa ja lääketieteessä.

Työssä esitellään pilvipalvelun määritelmä ja pilviominaisuuksien aiheuttamia pilvikohtaisia turvallisuusvaatimuksia sekä lyhyesti Euroopan Unionin tietoturva-asetus kuluttajien ja organisaatioiden kannalta.

Big datan käsittelyyn liittyvässä osiossa on esitelty Hadoop-alusta ja siihen liittyviä projekteja sekä R ja Python -ohjelmointikielet, joita käytetään datan analysoinnissa.

Osiossa, jossa käsitellään pilvipalvelutarjoajien ratkaisumalleja, esitellään joidenkin tarjoajien pilvessä toimivia big dataan liittyviä ratkaisuja. Liitteistä löytyy .CSV-tiedoston käsittelyä Google Dataprep -työkalulla.

2 BIG DATA KÄSITE

2.1 Mitä big data on?

Big data on suurien datatiedostojen kokoelma, joiden koot ylittävät yleisesti käytettyjen ohjelmistotyökalujen mahdollisuuden tallentaa, louhia, hallita ja käsitellä tietoja ”kohtuullisessa” ajassa. Big data edellyttää uutta tekniikkaa ja teknologiaa auttamaan monipuolisten, monimutkaisten ja suurten datamäärien käsittelyä. (Concessao 2016, 10.)

Hurwitz, Nugent, Halper & Kaufmanin (2013, 280) mukaan big data on mahdollisuus hallita suurta määrää erilaista tietoa, oikeaan nopeuteen ja oikeaan aikaan, jotta reaaliaikaisia analyysejä ja vastauksia voidaan saada. Suuri datamäärä jakautuu tyypillisesti kolmen ominaisuuden mukaan, jotka ovat määrä, nopeus ja vaihtelevuus.

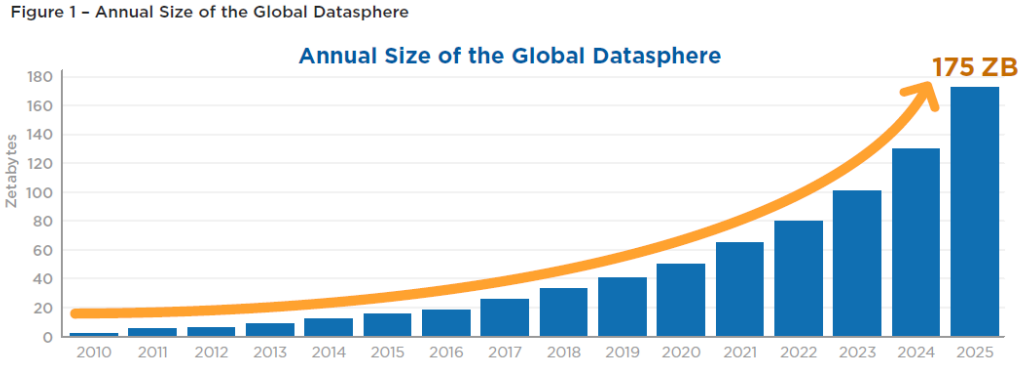
2.2 Big datan ominaisuudet

Doug Laney julkaisi työskennellessään META Group -yrityksessä vuonna 2001 raportin, jossa oli maininta datamäärien nopeasta kasvusta ja vaihtelevuudesta tulevina vuosina. Samassa raportissa olivat big dataan liittyvät alkuperäiset V-termit: Variety, Velocity ja Volume. Näillä kolmella V:llä kuvataan datan vaihtelevuutta, nopeutta ja määrää. Jo noin vuoden 2005 tienoilla alettiin puhua big datasta, mutta big datasta tuli ilmiö vasta vuonna 2011. (META Group 2001; Salo 2014, 26.)

2.3 Datan määrä (Volume)

Datan määrällä viitataan siihen kerättyyn ja varastoituun sähköiseen dataan, joka kasvaa lisääntyvässä määrin. Big datan tiedetään olevan suuri. Olisi helppoa määrätä tietty koko tarkoittamaan suurta tässä yhteydessä, mutta se, mitä pidettiin suurena kymmenen vuotta sitten, ei ole enää suuri tämän päivän mittapuulla. Datan keruu kasvaa sellaisessa määrin, että mikä tahansa valittu raja olisi vääjäämättömästi pian vanhentunut. (Holmes 2017, 16-17.)

IDC julkaisi vuonna 2018 työntekijöidensä David Reinselin, John Gantzin ja John Rydningin tekemän The Digitization of the World From Edge to Core -tutkimuksen, jossa haastateltiin 2400 yrityksen päätöksentekijää ja eri toimialueiden IT-johtajia. Nämä vastasivat organisaatioidensa kehittyneestä teknologiasta, kuten datan hallinnoinnista ja varastoinnista, IoT-laitteista ja koneoppimisesta. (IDC 2018, 1-6).



Kuvio 1. IDC:n tutkimukseen perustuva datan kehitys vuodesta 2010 vuoteen 2025. (IDC 2018, 6.)

2.4 Datan nopeus (Velocity)

Dataa tulee jatkuvasti eri lähteistä, kuten webistä, älypuhelimista ja erilaisista sensoreista. Nopeus liittyy määrään: mitä nopeammin dataa tuotetaan, sitä enemmän sitä on. Esimerkiksi sosiaalisen median viestejä toimitetaan ”lumipalloefektinä” ihmiseltä toiselle. Nämä viestit kulkevat nopeasti ympäri maailmaa. Nopeus viittaa myös datan sähköiseen käsittelynopeuteen. (Holmes 2017, 18.) Muun muassa F1-autoista saatava anturidata, joka on langattomasti lähetetty tallin insinööreille, täytyy analysoida reaaliaikaisesti, jotta kuljettaja saa välttämättömät ohjeet auton säätämiseksi. Vibrant Publishersin (2017, 8) mukaan nopeudella viitataan nopeuteen, jolla suurta datamäärää tuotetaan ja käsitellään.

Vaihtelevuutta voidaan pitää nopeuskäsitteen lisäulottuvuutena. Tällöin viitataan datavirtojen muuttuvaan nopeuteen, kuten datavirtojen kasvuun huippuaikoina. Tämä on merkittävää, koska tietokonejärjestelmät ovat alttiita virheille datavirtojen muutosten aikana. (Holmes 2017, 18.)

2.5 Datan vaihtelevuus (Variety)

Usein puhutaan internetistä ja webistä (World Wide Web) samana asiana, mutta ne ovat itseasiassa hyvin erilaisia. Internet on tietokoneverkkojen verkosto, joka sisältää tietokoneita, tietoverkkoja, paikallisia tietokoneverkkoja (LAN), satelliitteja, matkapuhelimia ja muita sähköisiä laitteita linkitettyinä toisiinsa ip-osoitteiden perusteella. (Holmes 2017, 17.)

Web-yhteydellä päästään käsiksi suureen määrään dataa lähteistä, osa on luotettavia ja osa epäilyttäviä, jolloin ollaan alttiina toistolle ja virheille. Perinteisten tilastojen vaatimaan selvään ja tarkkaan dataan on pitkä matka. Internetin palveluntarjoajien asiakkaat voivat ottaa yhteyden internetiin ja he pääsevät siten sisään webiin ja muihin palveluihin. (Holmes 2017, 17.)

Vaikka webistä kerätty data voi olla strukturoitua, strukturoimatonta tai semistrukturoitua, suurin osa webistä saadusta big datasta on strukturoimatonta. Muun muassa Twitterin käyttäjät julkaisevat noin 500 miljoonaa 140-merkkistä twiittiä päivässä maailmanlaajuisesti. Nämä lyhyet viestit ovat kaupallisesti arvokkaita ja ne analysoidaan sen mukaan, ovatko ne ilmaistu positiivisesti, negatiivisesti vai neutraalisti. Vaikka suuri datavalikoima voi olla sairaaloille, armeijalle ja moniin kaupallisiin tarkoituksiin kerättyä, se voi olla salaista strukturoituna, strukturoimattomana tai semistrukturoituna. (Holmes 2017, 17-18.)

Vibrant Publishersin (2017, 8) mukaan vaihtelevuus on tärkeää, koska pikaviestejä, sähköpostiviestejä, kuvia, äänitiedostoja, videoita, maantieteellistä dataa ja paljon muuta toimitetaan joka sekunti. Strukturoitua ja strukturoimatonta dataa täytyy prosessoida, analysoida ja varastoida. Kaikki tämä tieto on arvokasta ja auttaa liike-elämää ja hallituksia tekemään ratkaisevia päätöksiä.

2.6 Datan totuudenmukaisuus (Veracity)

Laneyn raportissa esittämiinsä kolmeen V:hen voidaan lisätä totuudenmukaisuus neljänneksi. Totuudenmukaisuudella viitataan kerätyn datan laatuun. Tilastollisten analyysien tunnuksena on ollut tarkka ja luotettava data. Data, jota tuotetaan digitaalisesti, on usein strukturoimatonta ja kerätty ilman kokeellista muotoilua tai käsitystä kysymyksistä, jotka herättäisivät mielenkiintoa. (Holmes 2017, 18-19.)

Datasta saatu tieto voi olla kyseenalaista. Esimerkiksi sosiaalisen median tuottama data voi olla luonteeltaan epätarkkaa ja epävarmaa ja usein informaatio ei yksinkertaisesti ole totta. Merkityksellisten tulosten saamiseksi muun muassa tilastoihin auttaa datan määrä. Kuitenkin on huomioitava, että suuri määrä dataa voi johtaa myös päinvastaiseen tulokseen väärien korrelaatioiden vuoksi. (Holmes 2017, 18-19.)

Vibrant Publishersin (2017, 8) mukaan totuudenmukaisuudella viitataan käsiteltävän datan luotettavuuteen.

McNeillin (2019) mukaan tietojen todenperäisyys on yleensä se, kuinka tarkka tai totuudenmukainen tietokokonaisuus on. Totuudenmukaisuus auttaa suodattamaan sen, mikä on tärkeää ja mikä ei, ja lopulta se luo syvemmän ymmärryksen datasta.

2.7 Datan visualisointi ja muut “V”:t

V-kirjain on tullut valintakirjaimeksi Laneyn kolmeen alkuperäiseen määritelmään, kuten myös pätevien määritysten lisäyksiin tai korvaamaan sellaiset termit, kuten datan haavoittuvuus (Vulnerability) ja datan käyttökelpoisuus (Viability), tärkeimpien lisäysten ehkä ollessa datan arvo (Value) ja datan visualisointi (Visualization). (Holmes 2017, 19).

Arvo viittaa yleensä big datan analyyseistä saatujen tulosten laatuun. Sitä on myös käytetty kuvaamaan kaupallisten yritysten myymiä datoja yrityksille, jotka käsittelevät niitä käyttäen omaa analytiikkaansa, joten termi viittaa myös datan liikemaailmaan. (Holmes 2017, 19.)

Van Rijmenamin (n.d.) mukaan kaikki data ei ole arvokasta. Arvolla viitataan siihen, miten organisaatiot käyttävät tietoja ja muuttavat organisaationsa informaatiokeskeiseksi yritykseksi, joka perustuu tietojen analyyseistä saatuun näkemykseen päätöksenteossaan.

Visualisointi ei ole big datan luonteenomainen piirre, mutta se on tärkeä analyyttisten tulosten esittämisessä ja tiedon välityksessä. Ympyrä- ja pylväskaavioita, jotka auttavat ymmärtämään pieniä datamääriä, on kehitetty auttamaan big datan visuaalista tulkintaa, mutta niiden soveltuvuudessa on rajoituksia. Kuvaajat tarjoavat monitahoisia esitystapoja, mutta ne ovat staattisia. (Holmes 2017, 19.)

Van Rijmenamin (n.d.) mukaan visualisointi tarkoittaa monimutkaisia kaavioita, joihin voi sisältyä monia datamuuttujia. Ne ovat kuitenkin ymmärrettäviä ja luettavissa. Visualisoinnit auttavat organisaatioita saamaan vastauksia haluttuihin kysymyksiin.

Big datan määrän ollessa jatkuvasti kasvamassa parhaat visualisoinnit ovat vuorovaikutuksessa käyttäjien kanssa ja säännöllisesti päivittyviä. Esimerkiksi automatkalla voidaan käyttää satelliittidataan perustuvaa GPS -paikannusjärjestelmää, joka on interaktiivinen, jatkuvasti päivittyvä järjestelmä sijainnin määritykseen. (Holmes 2017, 19-20.)

Big datan neljä pääominaisuutta, eli määrä, vaihtelevuus, nopeus ja totuudenmukaisuus, ovat merkittävä haaste big datan hallinnoinnissa. Hyödyt, jotka odotetaan saavutettavan kohtaamalla tämä haaste ja toivottavia vastauksia kysymyksiin, saadaan louhimalla big dataa. (Holmes 2017, 20.)

3 BIG DATAN LOUHINTA

Sanonta “data on uusi öljy” on yleistynyt teollisuudessa, liike-elämässä ja politiikassa. Sanonnalla tarkoitetaan, että data on öljyn lailla erittäin arvokasta, mutta sitä pitää käsitellä ennen kuin sen arvo realisoituu. Sanontaa käyttävät ensisijaisesti data-analytiikkatarjoajat taktiikkana tuotteidensa myymisessä vakuuttaessaan yhtiöille, että big data on tulevaisuus. Big datan vertaus öljyyn pitää paikkansa toistaiseksi. Kun löydetään öljysuoni, omistetaan myyntikelpoinen kauppatavara. Niin ei ole big datan kanssa. Ellei data ole oikeanlaista, sillä ei voi tuottaa mitään millä olisi arvoa. Ongelmia voi syntyä datan omistusoikeuksista ja yksityisyydestä. Toisin kuin öljy, data ei ole rajallinen resurssi. Big datan louhimisella tarkoitetaan hyödyllisen ja arvokkaan tiedon poimimista valtavasta datamäärästä. (Holmes 2017, 20.)

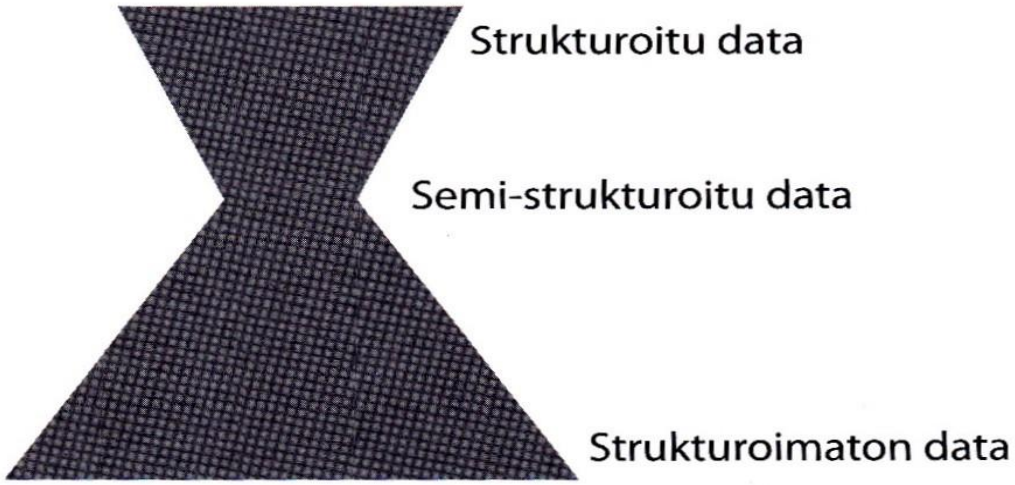
Kolb & Kolbin (2013, 27) mukaan data tulee arvokkaaksi vain silloin, kun siitä saadaan tietoa. Se tulee tiedoksi, kun siitä saadaan vastaukset kysymyksiin: kuka, mitä, missä, milloin ja kuinka?

Käyttämällä big datan louhintaa sekä koneoppimisen metodeja ja algoritmejä on mahdollista ennakoida ja huomata epätavallisia toistuvuuksia ja säännöttömyyksiä datassa. Tiedon saamiseksi suurista datamääristä voidaan käyttää valvottua tai valvomatonta koneoppimista. Valvottua koneoppimista voidaan verrata karkeasti ihmisen oppimiseen. Käyttämällä harjoitusdataa, jossa oikeat mallit on luokiteltu, tietokoneohjelma kehittää säännön tai algoritmin uusien mallien luokitteluun. Tämä algoritmi tarkistetaan käyttämällä testidataa. Valvomattomassa koneoppimisessa algoritmit käyttävät luokittelematonta syöttödataa, eikä mitään kohdetta ole annettu. Ne ovat suunniteltuja tutkimaan dataa ja löytämään piileviä rakenteita. (Holmes 2017, 20-21.)

Big datan manuaalinen käsittely ei ole käytännössä mahdollista, vaan siihen tarvitaan teknologiaa. Tähän ovat syynä datan liian suuri kirjavuus, muutosvauhti ja määrä. Periaatteessa big datan louhintaan liittyvät käytännöt ja prosessit ovat samoja kuin muussakin datan louhinnassa. Organisaatiot voivat joutua soveltamaan uutta teknologiaa ja tekniikoita omiin tarkoituksiinsa sopiviksi louhiessaan big dataa, datan monimuotoisuuden takia. (Salo 2013, 95.)

4 DATATYYPIT

Kaksi yleistä datatyyppiä ovat strukturoimaton ja strukturoitu data. Näiden välissä on semistrukturoitu data. (Salo 2013, 22.)



Kuva 1. Strukturoitu, semi-strukturoitu, strukturoimaton data. (Salo 2013, 22.)

Kuva 1 havainnollistaa datatyyppien suhteet toisiinsa. Strukturoimatonta eli rakenteetonta dataa on eniten, noin 80%. Strukturoitua eli rakenteellista dataa on noin 20%. Rakenteetonta dataa, joka lisääntyy huomattavasti, ei tallenneta ja analysoida riittävästi sen hintavuuden vuoksi, vaikka siitä voisi olla esimerkiksi markkinaetua liikeelämässä. (Salo 2013, 22, 25.)

Rakenteellisen ja rakenteettoman datan välillä on useita välimuotoja, joita kutsutaan semistrukturoiduksi dataksi. Rakenteettomasta datasta tulee rakenteellista, kun siihen lisätään tunnistetietoja. Dataa voidaan kutsua myös liikkuvaksi tai paikallaan pysyväksi dataksi. Liikkuvaa dataa on vaikeampi louhia ja hallita, koska sitä tuotetaan hyödynnettäväksi eri lähteistä, kuten erilaisista antureista, valvontajärjestelmistä ja IoT-laitteista, jotka tuottavat paljon liikkuvaa dataa. Big data koostuu näistä kaikista datatyypeistä. (Salo 2013, 22-25.)

Hurwitz ym. (2013, 26) mukaan termi strukturoitu data viittaa dataan, joka sisältää tunnistetietoja, kuten numeroita/lukuja, päivämääriä sekä sana- ja numeroryhmiä. Suurin osa asiantuntijoista on sitä mieltä, että rakenteellista dataa on noin 20% kaikesta datasta. Sitä varastoidaan yleensä relaatiotietokantoihin, joista voi tehdä kyselyjä käyttämällä strukturoitua kyselykieltä (SQL).

Rain (2019) mukaan strukturoitu data viittaa järjestäytyneisiin tietoihin, jotka ovat tallennettavissa ja käytettävissä tietokannoista yksinkertaisilla hakukonealgoritmeillä. Esimerkiksi yritysten tietokannoissa olevat taulukot, jotka sisältävät järjesteltyjä tietoja työntekijöistä, työtehtävistä ja palkoista ovat strukturoitua dataa.

Hurwitz ym. (2013, 29-30) mukaan strukturoimattomassa datassa ei ole rakennetta ja sitä on noin 80%. Rakenteetonta dataa on eniten ja se laajenee nopeasti. Termi strukturoimaton data saattaa johtaa harhaan, koska jokainen dokumentti voi sisältää oman erityisen rakenteensa tai dokumentin luomiseen käytetyn ohjelman muotoilun, kuitenkin dokumentin sisältö on strukturoimatonta, ellei se sisällä tunnisteita.

Rain (2019) mukaan strukturoimattomalla datalla viitataan dataan, jolla ei ole mitään tiettyä muotoa tai rakennetta. Rakenteettoman datan käsittely ja analysointi on vaikeaa ja aikaa vievää. Rakenteetonta dataa ovat esimerkiksi sähköpostit.

Hurwitz ym. (2013, 30) mukaan semistrukturoitu data on rakenteellisen ja rakenteettoman datan välissä. Semistrukturoitu data ei välttämättä mukaudu kiinteään eli rakenteelliseen kaavaan, mutta se voi olla itsestään kuvaileva ja voi sisältää yksinkertaisia nimiöitä tai arvopareja.

Rain (2019) mukaan semistrukturoitu data sisältää rakenteellista ja rakenteetonta dataa ja viittaa tietoihin, joita ei ole luokiteltu tiettyyn tietokantaan. Se voi sisältää kuitenkin tietoja ja tunnisteita, jotka erottavat yksittäiset elementit datassa.

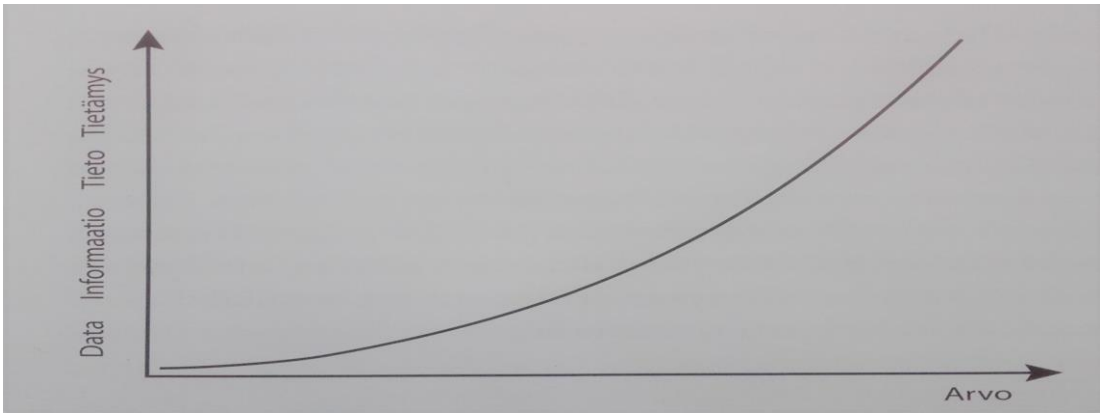
4.1 Datan analysointi

Pickellin (2018) mukaan kuvaileva (Descriptive) analyysi luo yksinkertaisia raportteja, kaavioita ja muita visualisointeja, jotka auttavat ymmärtämään, mitä tietyssä kohdassa tapahtui. Kuvaileva analyysi koskee vain aiempia tapahtumia. Diagnostinen (Diagnostic) analyysi taasen antaa syvemmän käsityksen tiettyyn ongelmaan, kun kuvaileva analyysi on enemmän yleiskatsaus. Yritykset voivat käyttää diagnostista analyysia ongelman ymmärtämiseen. Tämä analyysi voi sisältää tekoälyn (AI) ja koneoppimisen näkökohtia. Ennustavassa (Predictive) analyysissa voidaan ennustaa, mitä seuraavaksi tapahtuu, kun yhdistellään kehittyneitä algoritmejä tekoälyn ja koneoppimisen kanssa. Ohjeellinen (Prescriptive) analyysi on äärimmäisen monimutkainen. Muiden analyyttisten työkalujen avulla voidaan tehdä omia päätelmiä, mutta ohjeellinen analyysi antaa todelliset vastaukset. Näihin raportteihin tarvitaan korkean tason koneoppimista.

Davisin (2019) mukaan kuvaileva analyysi tutkii datan historiaa ja vastaa kysymyksiin siitä, mitä on aiemmin tapahtunut. Diagnostinen analyysi tunnistaa rakenteita, sillä se etsii datansisäisiä yhtäläisyyksiä ja vastaa kysymykseen, miksi jotakin tapahtui. Ennustava analyysi käyttää nykyistä ja aikaisempaa dataa tulevan toiminnan ennustamiseen ja vastaa kysymykseen, mitä tulee tapahtumaan. Ohjeellinen analyysi soveltaa sääntöjä ja mallinnusta parempien päätösten tekoa varten ja vastaa kysymykseen, mitä tehdään.

5 DATASTA TIETÄMYKSEEN

Big datan datamäärät koostuvat eri lähteistä muodostuneista datamassoista. Niistä halutaan kehittää informaatiota, joka on perustana tietämykselle ja tiedolle päätösten tekemiseen. (Salo 2014, 32.)



Kuva 2. Datan, informaation, tiedon ja tietämyksen keskinäinen suhde big datassa. (Salo 2013, 27.)

Kuva 2 havainnollistaa, miten datasta saadusta informaatiosta voidaan analysoida pitkällä aikavälillä tietoa ja tietämystä ja saada viitteitä tulevaisuuteen. Big data -työkaluilla louhitaan data ja informaatio data-analyytikoille, joiden tehtävänä on analysoida saatu data arvokkaaksi tiedoksi ja tietämykseksi. (Salo 2014, 32-33.)

Kolb ym. (2013, 26-27) mukaan data koostuu mistä tahansa binäärisestä syötöstä, jota voidaan prosessoida tietokoneella. Prosessista saadaan tietoa ymmärtämällä datan varsinainen säännönmukaisuus ja yhtäläisyydet. Tietämys syntyy, kun tieto mahdollistaa datan ennustettavuuden ja käytön tuottamaan positiivisia lopputuloksia. Sen sijaan, että tiedetään vain mitä ja missä, voidaan tunnistaa, miksi yksittäinen asia tapahtuu ja hyödyntää se.

Vaikka tietoa ja tietämystä voitaisiinkin tuottaa automaattisesti, tarvitaan kuitenkin ihmisen tietotaitoa niiden saamiseksi ymmärrettävimmiksi, esimerkiksi yritysten johtoryhmille. Työvälineiden valinnoilla voidaan saada tallennettava pysyvä data ja liikkuva data laajemmiksi, jolloin saadaan enemmän informaatiota jalostettavaksi. (Salo 2014, 32-33.)

6 BIG DATAA HYÖDYNTÄVIÄ ALUEITA

Big dataa hyödynnetään useilla eri alueilla. Tässä osiossa kerrotaan, miten se voi auttaa kehittämään julkishallintoa, liike-elämää ja terveydenhuoltoa. Myös pankki-, vakuutus-, media- ja teollisuusala hyödyntävät big dataa toiminnoissaan.

6.1 Julkishallinto

Julkisen sektorin data on osittain avointa, koska sen tehtävänä on palvelu. Julkishallinnossa avoimen datan jakaminen vastikkeetta auttaa läpinäkyvään hallintoon ja tehokkaaseen demokratiaan saman datan ollessa käytössä kaikilla osapuolilla. Uusia palveluita voidaan kehittää julkisesta data-aineistosta, jonka pitää olla monipuolista, pääsyn teknisesti yksinkertaista ja standardoitua sekä laajaa. Julkishallinnosta saatavasta datasta voidaan yritysten kehittämien innovaatioiden avulla avata uusia rajapintoja palveluihin, jotka parantavat työllisyyttä ja lisäävät kulutusta sekä tukevat kannattavaa liiketoimintaa tuottamaan verotuloja. Kun julkishallinnon sisäinen tehokkuus kasvaa, avoimen datan jakaminen auttaa toimintojen keskittämiseen ja päällekkäisyyksien välttämiseen sekä erikoistumisasteen paranemiseen. Julkishallinnossa, kuten avoimessa liiketoiminnassakin, ovat innovaatiot mahdollisia. (Salo 2013, 32,35.)

6.2 Liike-elämä

Yrityksissä kerätystä datasta voidaan analysoimalla saada selvyys asiakasmääristä ja asiakkaiden kulutustottumuksista. Tällä analysoidulla datalla saadaan myös kokonaiskuva tehtävistä parannuksista liiketoimintaan, asiakasrajapintoihin ja tuotekehittelyyn. (Salo 2013, 33.)

Esimerkki 1. Dataa kerätään yritysten verkkosivuilta, yrityksissä olevista päätelaitteista ja markkinatutkimuksilla. Henkilöstö voi tehdä asiakastyöskentelyn ohella havaintoja, joista kerätty data voidaan analysoida tuoton parantamiseksi. (Salo 2013, 33-34.)

Esimerkki 2. Kuluttajan ostaessa tuotteita yrityksen verkkokaupasta, samoin kuin fyysisestä asioinnista myymälässä kuluttajan käyttäessä kanta-asiakaskorttia, kertyy tietoja asiakasrekisteriin ostoista ja maksutavoista. Asiakasrekisterien tietojen pohjalta voidaan kohdentaa mainontaa ja palvelujen tarjontaa vastaamaan kuluttajien tarpeita. Kuluttajat tuottavat dataa itsestään myös käyttämällä esimerkiksi Googlen hakukonetta. Useiden kuluttajien hakuja yhdistelemällä ja analysoimalla voidaan saada ajankohtaista tietoa kuluttajien mielenkiinnon kohteista, joista voidaan ennustaa ja mallintaa kaupallista arvoa sisältävää tietoa. (Salo 2013, 40-42.)

6.3 Terveydenhuolto

Terveydenhuoltoala tuottaa potilaista suuria määriä tietoja, joiden saatavuus, hallinta ja tulkinta ovat ratkaisevan tärkeitä paremman ja tehokkaamman hoidon saamiseksi. Paremmat käsitykset hoidosta, tutkimuksesta ja tehokkaasta käytännöstä edistävät big data -tietojen tarvetta terveydenhuollossa. (Patidar 2018.)

Esimerkki 1. Syövän hoidossa big datasta saatavia tietoja käytetään etsittäessä yksittäisten syöpäsolujen tietojen malleja ja geneettisten biomerkitsimien löytämiseksi. Yhteisten piirteiden löytäminen voi auttaa ennustamaan, miten yksittäiset kasvaimet voivat muuttua ja mikä lääkehoito olisi tehokkain. Potilaan sairaushistoriaa ja DNA-tietoja voidaan käyttää määrittelemään paras hoito potilaille, joilla on samanlainen sairausmalli ja geneettisyys. Laajempia väestötietoja voidaan analysoida, jotta potilaiden hoitostrategiat saadaan erilaisten elämäntapojen, maantieteellisten ja syöpätyyppien perusteella. Big datasta saadun tiedon käyttämisen päätavoite syöpähoidoissa on räätälöity lääkehoito ja vakavien sivuvaikutusten estäminen. (Bayer AG, 2018.)

Esimerkki 2. Big datan tiedot auttavat ennustamaan muun muassa ebolaepidemian leviämistä. Väestön liikkumista voidaan seurata esimerkiksi matkapuhelimien sijaintitietojen avulla, jolloin voidaan ennakoida viruksen leviäminen. Näin saadaan tietoa epidemian vaikutusalueista ja voidaan suunnitella hoitokeskuksia sekä rajoittaa liikkumista näillä alueilla. (Patidar 2018.)

7 IOT-LAITTEET JA ROBOTIT

IoT-laitteita ja robotteja kehitetään automatisoimaan toimintoja eri tarkoituksiin. Big datasta kehitettyjä tekoälytekniikoita ja IoT-toteutuksia hyödynnetään esimerkiksi kulkuneuvoissa, älykodeissa ja älykaupungeissa. Robotteja käytetään muun muassa lääketieteessä, maataloudessa ja teollisuudessa, joista on muutama esimerkki tässä osiossa.

7.1 Kulkuneuvot

Techopedia Inc. (n.d.) mukaan itseohjautuva auto tunnetaan myös nimellä kuljettajaton auto, autonominen auto tai robottiajoneuvo. Siinä käytetään erilaisia tekniikoita, kuten antureita ja muita laitteita esimerkiksi törmäysten välttämiseksi. Siihen voi sisältyä myös GPS-antureita navigoinnin helpottamiseksi.

NVIDIA Corporationin (2016) mukaan itseohjautuvat autot tulivat mahdollisiksi, kun otettiin käyttöön syväksi oppimiseksi kutsuttu tekoälytekniikka. Syvä oppiminen on välttämätöntä itsenäisille ajoneuvoille, koska kukaan ei voi kirjoittaa ohjelmistoja, jotka ennakoivat kaikki mahdolliset skenaariot, joita itseohjautuva auto voi kohdata. Syvällä oppimisella auton tietokone voi oppia, mukautua ja kehittyä. Autonomisilla autoilla voidaan päästä parempaan tulevaisuuteen. Niillä on mahdollisuus vähentää onnettomuuksia ja hiilidioksidipäästöjä sekä tarjota liikkumismahdollisuuden ihmisille, jotka eivät voi ajaa.

Vuonna 2016 Amazon ilmoitti tehneensä ensimmäisen kaupallisen GPS-paikannusjärjestelmää käyttävän droonin. Vaikka matkustajien kuljettamiseen drooneilla on vielä pitkä matka, käytetään niitä nykyään muun muassa viljan ruiskutuksiin maatiloilla ja sotilaallisiin tarkoituksiin. Itseohjautuvien kulkuvälineiden kehitys, autoista lentokoneisiin, näyttää vääjäämättömältä. Älykkäät kulkuneuvot ovat kehityksen alussa yleiseen käyttöön, mutta älylaitteet ovat jo osa nykyaikaista kotia. (Holmes 2017, 107-108.)

7.2 Älykodit

Holmesin (2017, 108) mukaan mitkä tahansa elektroniset laitteet, jotka voidaan asentaa kotiin ja hallita etäyhteydellä, kuten älytelevisiot, älypuhelimet ja tietokoneet, ovat älylaitteita ja siten IoT:n osia. Älykodissa voi olla puheella ohjattu keskushallintajärjestelmä, jolla säädetään valaistusta, lämmitystä, autotallin ovia ja muita kodin laitteita.

Comcastin (2019) mukaan älykoti on koti, joka on varustettu internet-yhteyteen kytketyillä laitteilla, joita asukkaat voivat kauko-ohjata tietokoneen, tabletin tai älypuhelimen kautta mistä ja milloin tahansa. Laitteet auttavat kodin tehtävien automatisoinnissa. Älykodissa laitteet voivat ”puhua” toisilleen suorittaakseen tiettyjä tehtäviä. Esimerkiksi, kun ovitunnistin havaitsee liikettä, se voi kytkeä valot päälle.

7.3 Älykaupungit

Holmesin (2017, 110-111) mukaan älykaupunkien teknologiaa ohjaavat erilliset, mutta kootut, IoT-toteutukset ja big data -hallintatekniikat. Esimerkiksi itseohjautuvat autot, terveyden etäseuranta ja älykodit kuuluvat älykaupungin ominaisuuksiin. Älykaupunki on riippuvainen big datasta, jota kerätään kaupunkiin asennetuista useista sensoreista. Nämä sensorit tuottavat suuren määrän dataa, jota täytyy valvoa ja analysoida keskustietokoneella tosiaikaisesti. Yhteisöllä voi olla esimerkiksi älykäs energiajärjestelmä, jolla voidaan ohjata katuvalaistusta, valvoa liikennettä ja seurata jätehuoltoa.

Gemalton (2019) mukaan älykkäässä kaupungissa on IoT-laitteiden verkko, joka siirtää dataa langattoman tekniikan ja pilven avulla. Pilvipohjaiset internet-sovellukset vastaanottavat ja hallitsevat tietoja reaaliajassa. Kaupunkien ekosysteemeihin voidaan olla yhteydessä älypuhelimien, mobiililaitteiden samoin kuin siihen kytkettyjen autojen ja kotien avulla. Laitteiden ja datan yhdistäminen kaupungin fyysiseen infrastruktuuriin ja palveluihin voi vähentää kustannuksia. IoT-laitteiden avulla voidaan parantaa energiajakelua, vähentää liikenneruuhkia ja jopa parantaa ilmanlaatua.

7.4 Robotit lääketieteessä

Esimerkki 1. DaVinci robottia käyttämällä leikkaukset voidaan tehdä vain muutamalla pienellä viillolla ja erittäin tarkasti, mikä tarkoittaa vähemmän verenvuotoa, nopeampaa paranemista ja vähentää infektioriskiä. Laite on kirurgin täydessä valvonnassa. Suuret teknologiayritykset kehittävät DaVincin mukaisia järjestelmiä, joilla on autonomisemmat ominaisuudet ja laajempi toimintakyky. (Tomlinson 2018.)

Esimerkki 2. Tiedemiehet, jotka toimivat yhteistyössä Cambridgen ja Aberystwythin yliopistojen kanssa kehittivät robottitutkijan ”Adam”, joka on onnistuneesti muotoillut ja testannut uusia perinnöllisyystieteen hypoteeseja. Nämä ovat johtaneet uusiin tieteellisiin löytöihin. (Holmes 2017, 106.)

Esimerkki 3. Sama tiedemiesryhmä, joka kehitti ”Adamin”, on kehittänyt Manchesterin yliopistossa robotin ”Eve”, jonka tarkoituksena on nopeuttaa lääkkeiden löytämisprosessia ja tekemään siitä taloudellisempaa. Tiedemiesryhmä on osoittanut tekoälyn potentiaalin käyttämällä ”Eveä” selvittämään voidaanko yhdistettä, jolla on syövän vastaisia ominaisuuksia, käyttää myös malariaa vastaan. (University of Cambridge 2015.)

7.5 Robotit maataloudessa

Maatalousrobotit ovat suunnitellut tekemään maatiloilla olevia erilaisia töitä. Vaikka robotit vähentävät työpaikkoja maataloudessa, samalla ne luovat uusia mahdollisuuksia automatisoitujen järjestelmien ohjelmointiin ja toimintaan. Maatalouden toimintaa kehittäviä IoT-yrityksiä ovat muun muassa IBM, FarmersEdge ja Cattle Watch. (Schmidt 2018.)

7.6 Robotit autoteollisuudessa

Autoteollisuudessa robotteja käytetään hitsauksessa, kokoonpanossa, tuotantokoneiden hoitamisessa, materiaalien käsittelyssä, osien siirtämisessä ja maalauksessa. (Acieta LLC. 2019).

Autonomiset robotit muuttavat tapoja ajoneuvojen suunnitteluun ja testaamiseen. Esimerkiksi Yamahan MOTOBOT, humanoidirobotti, voi ajaa ja testata moottoripyörää tuottaen reaaliaikaista dataa entistä tarkemmin. MOTOBOT voi olla käytössä myös muiden ajoneuvojen testaamisessa. (Schmidt 2018.)

8 PILVIPALVELUT

Pilvipalveluilla mahdollistetaan suuri tallennustila ja laskentateho big datan käsittelyyn. Pilvipalveluilla on useita eri ominaisuuksia, palvelumalleja sekä käyttöönottomalleja, jotka mahdollistavat pilvipalveluiden soveltuvuuden eri käyttötarkoituksiin. Pilvipalveluiden käsitteestä ja yleisestä määritelmästä kerrotaan seuraavissa kappaleissa tarkemmin.

8.1 Pilvipalvelun käsite

Salon (2013, 103) mukaan big data -pilvipalvelu on palvelu, jota aiemmin markkinoitiin pelkällä pilvipalvelu -termillä.

Holmesin (2017, 36) mukaan pilvipalveluilla viitataan toisiinsa yhteydessä olevien serverien verkostoon, jota hallitaan datakeskuksissa ympäri maailman. Nämä datakeskukset mahdollistavat big datan keskitetyn varastoinnin. Eri yritysten tarjoamilla pilvipalveluilla voidaan internetin kautta käyttää etäservereitä, varastoida sekä hallita tiedostoja ja käyttää sovelluksia. Niin kauan kuin tietokoneessa tai muussa laitteessa on tarvittava ohjelma, jolla päästään pilvipalveluun, voidaan analysoida tietoja milloin ja missä tahansa sekä antaa myös muille valtuudet tehdä se. Lisäksi voidaan myös käyttää ohjelmistoa, joka on pilvipalvelussa mieluummin kuin omassa laitteessa. Tietojenkäsittely pilvessä (Cloud computing) ei ole vain pääsy internetiin, vaan vaatii myös oikeanlaiset välineet tietojen varastointiin ja prosessointiin.

8.2 NIST:n määritelmä pilvipalveluille

Yhdysvaltalainen NIST (National Institute of Standars and Technology) julkaisi vuonna 2011 tietojenkäsittelytieteentutkijoiden Peter Mellin ja Timothy Grancen yleisesti käytössä olevan pilvipalveluiden määritelmän: Cloud computing on pilvipalvelumalli, jossa verkkoyhteydellä mahdollistetaan pääsy pilvessä oleviin muunneltaviin tietovarantoihin, kuten verkkoyhteyksiin, palvelimiin, tietovarastoihin, sovelluksiin ja

palveluihin. Niitä voidaan ottaa käyttöön ja poistaa käytöstä itse nopeasti pienellä hallinnointivaivalla tai vuorovaikutuksessa palveluntarjoajaan. Määritelmässä listataan pilvipalvelun viisi keskeistä ominaisuutta: palveluntarjoajalta hankittu itsepalvelu (on-demand self-service), pääsy useilla päätelaitteilla palveluihin (broad network access), resurssivarannot (resource pooling), kapasiteetin nopea joustavuus tai laajennus (rapid elasticity or expansion) ja tarkka käytön mittaaminen automaattisesti (measured services). (NIST 2011, iii-2.)

Määritelmässä listataan myös kolme palvelumallia: ohjelmisto-, alusta- ja infrastruktuuripalvelumallit. Ohjelmistopalvelu (SaaS) on palvelu, jossa asiakas käyttää palveluntarjoajan tarjoamia pilvipalvelussa toimivia sovelluksia. Alustapalvelu (PaaS) on palvelu, jossa palveluntarjoaja tarjoaa sovelluskehitysympäristön ja apuohjelmia. Näiden lisäksi asiakas voi kehittää omia sovelluksia alustan päälle. Infrastruktuuripalvelu (IaaS) on palvelu, jossa palveluntarjoaja tarjoaa asiakkaalle tietokoneiden verkkoyhteyksiä, laskentatehoa ja tallennustilaa. Käyttöjärjestelmistä lähtien asiakas voi itse toteuttaa tai valita loogiset yhteydet ja ohjelmistot. (NIST 2011, 2-3.)

Määritelmässä on neljä käyttöönottomallia: yksityinen pilvi, yhteisöllinen pilvi, julkinen pilvi ja hybridipilvi. Yksityinen pilvipalvelu on tarkoitettu yksinomaan yhdelle organisaatiolle, joka koostuu useista kuluttajista. Se voi olla organisaation, kolmannen osapuolen tai niiden yhdistelmän omistuksessa ja hallinnassa. Yhteisöllinen pilvi-infrastruktuuri on tarkoitettu ennalta rajatuille käyttäjäyhteisöille organisaatioissa, joissa yhteisöillä on samoja turvallisuusvaatimuksia ja tavoitteita pilvipalvelun käytössä. Se voi olla yhden, useamman yhteisön, kolmannen osapuolen tai niiden yhdistelmän omistuksessa ja hallinnassa. Julkinen pilvipalvelu on tarkoitettu avoimeen käyttöön, se voi olla yrityksen, yliopiston, hallituksen organisaation tai niiden yhdistelmän hallinnassa ja omistuksessa. Hybridipilvi on kahden tai useamman pilvipalvelun yhdistelmä, jossa käytetään standardisoituja rajapintoja. NIST:n määritelmä on tarkoitettu tueksi pilvipalveluiden ja käyttöönottostrategioiden laajaan vertailuun. (NIST 2011, 1, 3.)

Pilvipalveluiden joustavuuden vuoksi analytiikkapalveluiden ja big datan tallennuspalveluiden käytön etuina katsotaan olevan, ettei etukäteen tarvitse olla tietoa tarvittavasta kapasiteetista, eikä tarvinne hankkia ohjelmistoja, laitteita ja tehdä pitkäaikaisia

sopimuksia. Pilvipalvelut ovat hintatasoltaan edullisia suurta tallennustilaa tarvitseville organisaatioille. (Salo 2013, 103-104.)

9 TIETOTURVA PILVIPALVELUISSA

NIST:n (2015, 10) mukaan monet big dataan suunnitellut järjestelmät toteutetaan pilviarkkitehtuurin avulla. Kaikissa strategioissa, joilla saavutetaan asianmukainen pääsynvalvonta ja tietoturvariskien hallinta big data -pilven ekosysteemiin yritysarkkitehtuurissa, on käsiteltäviä pilviominaisuuksien aiheuttamia pilvikohtaisia turvallisuusvaatimuksia. Näitä vaatimuksia ovat laaja verkkoyhteys, kuluttajien näkyvyyden ja valvonnan väheneminen, järjestelmän dynaamiset rajat sekä kuluttajien palveluntarjoajien roolit ja vastuut, palvelun vuokra-aika, tietokannat, mitatut palvelut, laajennus pyydettäessä, joustavuus, hinnoittelun optimointi, automatisointi ja virtualisointi. Nämä pilviominaisuudet aiheuttavat usein erilaisia turvallisuusriskejä kuin perinteiset IT-ratkaisut. Ne muuttavat organisaatioiden tietoturvanäkemystä. Turvallisuuden säilyttämiseksi, kun tietoja siirretään pilveen, organisaatioiden on etukäteen tunnistettava kaikki pilvikohtaiset, riskiin mukautetut turvatarkastukset tai komponentit. Joissakin tilanteissa voi olla tarpeen pyytää pilvipalveluiden tarjoajilta sopimus- ja palvelutason mukaisesti, että kaikki turvakomponentit ja -ohjaukset toteutetaan täysimääräisesti ja tarkasti.

Holmesin (2017, 91, 94) mukaan epätodennäköistä on, jos kaikki tieto on tallennettuna pilveen, että kaikki tallennettu tieto nykypäivän pitkälle kehittyneistä järjestelmistä katoaisi. Toisaalta, jos halutaan poistaa jotakin tietoa, on luotettava, että palveluntarjoaja poistaa kaikki kopiot. Toinen tärkeä seikka on, että kontrolloidaan, kenellä on pääsy pilveen tallennettuun dataan. Jos halutaan saada big data turvallisemmaksi, datan salaus on tärkeää. Verkkoturvallisuuteen vaikuttaa palomuuri, joka eristää ulkopuolisen luvattoman pääsyn dataan internetin kautta. Vaikka verkkosivu olisi suojattu suoralta hyökkäykseltä, varsinkin salaamaton data on alttiina viruksille ja troijalaisille. Verkkourkinnassa lähetetään haitallista koodia sisältävä exe-tiedosto tai kysellään henkilökohtaisia tietoja, kuten salasanoja sähköpostitse. Big dataa koskeva pääongelma on kuitenkin hakkerointi.

10 EUROOPAN UNIONIN TIETOSUOJA-ASETUS

Big dataa tuottavat muun muassa sähköpostit, pikaviestit ja älylaitteet ympäri maailman. Nämä voivat sisältää henkilötietoja, joiden käsittelyyn Euroopan Unionin alueelle on säädetty yleinen tietosuoja-asetus (General Data Protection Regulation, GDPR). Tietosuoja-asetus parantaa kuluttajasuojaa ja määrittelee henkilötietoja käsittelevien organisaatioiden vastuut ja velvollisuudet tietojenkäsittelyyn.

10.1 GDPR ja kuluttajat

Waden (2018) mukaan EU:n yleisen tietosuoja-asetuksen (GDPR) pyrkimyksenä on suojella kansalaistensa yksityisyyttä. Kaikkien jäsenvaltioiden on noudatettava asetusta. GDPR:n alaisuuteen kuuluvat kaikki, joilla on asiakkaita EU:n alueella ja työskentelevät tietojenkäsittelyn parissa. Kuluttajien kannalta positiivista on ei-toivottujen mainosten väheneminen ja pienentynyt riski henkilökohtaisten tietojen joutumisesta vääriin käsiin. Kuluttajat voivat paremmin muuttaa tai poistaa itseään koskevia vääriä tietoja. Negatiivisena puolena on yksilöllisen palvelun saamisen hankaluus.

Scottin (2018) mukaan GDPR parantaa yksityisyyden suojaa. Kuluttajien kannalta positiivista on, että yritysten on lähetettävä selkeät suostumuspyynnöt sähköpostimarkkinointia varten. Vastauksen puute ei tarkoita suostumusta. Kuluttajat voivat pyytää tietojensa poistamista, päivittämistä tai muuttamista, jos ne ovat virheellisiä tai tarpeettomia.

10.2 GDPR ja organisaatiot

Waden (2018) mukaan GDPR aiheuttaa muutoksia organisaatioiden IT-järjestelmiin ja toimintatapoihin. Se tarkoittaa, että usein vanhojen prosessien mukautus ei riitä, vaan on rakennettava uusia prosesseja ja järjestelmiä tai suunniteltava uudelleen olemassa olevat järjestelmät.

Carsonin (2018) mukaan muutos on kallista kaikille, joilla on toimintaa EU:n alueella. IAPP-EY arvioi vuoden 2017 yksityisyyden hallintaa koskevassa raportissa, että Fortune Global 500 -yritykset käyttävät noin 7,8 miljardia dollaria GDPR-vaatimusten mukaisuuteen henkilöstön palkkaamisesta aina tuotteiden ja palveluiden muuttamiseen.

11 TYÖKALUJA BIG DATAN KÄSITTELYYN

11.1 R

R -kieli on avoimen lähdekoodin projekti, jonka kehitti John Chambers kollegoineen työskennellessään Bell laboratoriossa. R sisältää ohjelmistopaketteja tietojen käsittelyyn, laskentaan ja graafiseen esitykseen. Siihen sisältyvät tehokkaat ominaisuudet tietojenkäsittelyyn- ja varastointiin, operaattorijoukko taulukoiden ja matriisien laskentaan sekä laaja työkalukokoelma data-analyyseihin. Analysoidut tiedot voidaan esittää graafisesti joko näytöllä tai paperilla. R on kehittynyt, yksinkertainen ja tehokas ohjelmointikieli, joka sisältää ehtoja, silmukoita, käyttäjän määrittelemiä toistokelpoisia toimintoja sekä syöttö- ja tulostemahdollisuudet. R on laajennettavissa pakettien avulla. R-jakelu sisältää noin kahdeksan pakettia ja lisää paketteja on saatavilla CRAN-palvelimilta, joilta löytyy laaja valikoima paketteja nykyaikaiseen tilastointiin. (The R Foundation.)

11.2 Python

Python on ilmainen avoimen lähdekoodin ohjelmointikieli, jota käytetään laajalti tieteellisessä laskennassa, akateemisessa maailmassa ja teollisuudessa. Se sisältää useita hyödyllisiä ja hyvin testattuja analytiikkakirjastoja, jotka sisältävät paketteja numeraaliseen laskentaan, data-analyysejä, tilastollisia analyysejä, visualisointia ja koneoppimista. (Madsen, Cormier, Von Stecher, Liu, Voronov, Gu & Wiley 2014.)

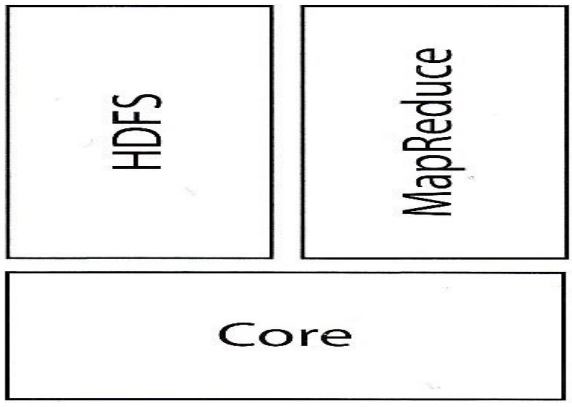
Python tukee olio-ohjelmointia ja kehittyneitä tietorakenteita, kuten listoja, sarjoja ja hakemistoja. NumPy-kirjaston avulla voidaan suorittaa matriisioperaatioita ja pandas-paketti tukee datakehyksiä. Näiden käyttöönotto helpottaa ja nopeuttaa datatoimintoja. (Madsen ym. 2014.)

R ja Python ovat hyödyllisiä ohjelmointikieliä data-analytiikassa. R:ssä on tilastollisien analyysien kattava tuki. Python integroituu paremmin eri järjestelmiin, reaaliaikaisiin tietovirtoihin sekä ohjelmistotyökaluihin mukaan lukien R, joten R:n ja Pythonin yhdistelmä tarjoaa hyvän työkalusarjan datan tutkijoille. (Madsen ym. 2014.)

11.3 Hadoop

Hadoop on Apache Foundationin avoimen lähdekoodin hyvin skaalautuva ja tehokas alusta, joka pystyy käsittelemään suuria määriä monimuotoista dataa hajautetussa klusteroidussa ympäristössä. Hadoopin sisältämät tärkeimmät ydinosat ovat Hadoop Distributed File System (HDFS) ja MapReduce-ohjelmointimalli. (Kaushik 2016.)

Vibrant Publishersin (2017, 12) mukaan Hadoop ei ainoastaan auta käsittelemään rakenteellista ja rakenteetonta dataa, joka on monimutkaista käsitellä, vaan auttaa myös analysoimaan tietoja, jotka ovat arvokkaita monilla toimialoilla, kuten terveydenhoito, liike-elämä ja koulutus.



Kuva 3. Hadoopin kolme ydinprojektia. (Salo 2013, 82.)

Kuva 3 havainnollistaa Hadoopin ydinosat. HDFS on hyvin skaalautuva ja suuren kapasiteetin tarjoava hajautettu tiedostojärjestelmä. Datalohkot replikoidaan ja tallennetaan hajautetusti klusteroidussa ympäristössä. MapReduce on ohjelmointimalli hajautetussa ympäristössä olevan suuren datamäärän rinnakkaiseen käsittelyyn. Map-osio suorittaa suodatuksen ja lajittelun, Reduce-osio tekee yhteenvedon Map-osion tiedoista. Hadoop 2-versiosta lähtien on ollut saatavilla YARN-teknologiaan (Yet Another Resource Negotiator) perustuva MapReduce-malli. Yarn on kehys laskennallisten resurssien tulosten prosessointiin. Yarn tarjoaa resursseja sovellusten suorittamiseen ja HDFS tarjoaa runsaasti tarvittavaa tallennustilaa. Hadoop Common sisältää yleisiä

työkaluja ja apuohjelmien kirjaston Hadoop-modulien tueksi. Sitä käytetään pääasiassa sovelluskehityksen aikana. (Kaushik 2016; The Apache Software Foundation 2018; Vibrant Publishers 2017, 30.)

11.4 Hadoopiin liittyviä projekteja

Apache Pig on ohjelmointityökalu, jota käytetään luomaan ohjelmia, jotka toimivat Hadoopissa käyttämällä Big Latin -ohjelmointikieltä. Se käyttää MapReduce-, HDFS- ja UDF-tiedostoja luomaan ohjelmia, jotka ovat vuorovaikutuksessa tietokannan kanssa ja käsittelevät tietoja. Se on avoimen lähdekoodin ohjelmointikieli, joka kehitettiin suurten tietomäärien käyttämiseen ja käsittelyyn. Big Latin hallinnoi datavirtaa, UDF:t hallinnoivat liitoksia, suodattimia, lukemista ja kirjoittamista. MapReduce ja HDFS hallinnoivat tietokannan käyttöä ja tallennusta. Apache Pig on hyvä suurille rakenteettomille tietoryhmille. Se tallentaa tietoja, milloin tahansa limittämällä toimintoja siten, ettei tarvitse käyttää tietokantaa. (Vibrant Publishers 2017, 63-64.)

Impala on avoimen lähdekoodin SQL-kyselymoottori. Impalalla saadaan nopeita ja interaktiivisia tuloksia jopa suurista tietomääristä. Impala on nopeampi kuin MapReduce tai joku muu Hadoopissa käytettävä SQL-moottori. Se tukee SQL-tyyppisiä kyselyitä tietojen käyttämiseen ja tiedostojärjestelmiä, kuten HDFS, Amazon S3 ja Apache HBase sekä monia tiedostomuotoja. Se voidaan myös integroida liiketoimintatietovälineisiin. (Vibrant Publishers 2017, 69-70.)

HBase on Hadoopin oma NoSQL-tietokanta, joka toimii HDFS:ssä ja tukee jaettua käsittelyä. Se käsittelee reaaliaikaista dataa avainarvoparina ja yhdistää MapReducen analyyttiset ominaisuudet. HBasella on nopea ja suora pääsy HDFS-tietokantaan tallennettuun tietomäärään. Se hakee nopeammin Hadoop-järjestelmään tallennettuja irtotietoja ja lähettää massatietoja HDFS-järjestelmään. HBase koostuu taulukoista, jotka ovat sarakeperheiden (Column Families) kokoelma. Jokainen sarakeperhe sisältää joukon niihin liittyviä sarakkeita, jotka sisältävät dataa avainarvopareissa. Koska HBase on sarakekeskeinen, tietyn sarakkeen perheen avaimen rivit pidetään yhdessä. Jokainen solu rivin, sarakepohjan ja sarakkeen yhdistelmässä sisältää aikaleiman, joka ilmaisee koska data on kirjoitettu tai version datasta. HBasen aikaleima tai versiointi

mahdollistaa käyttäjän pääsyn mihin tahansa versioon tallennetusta datasta. (Vibrant Publishers 2017, 106-107.)

Apache Flume on hajautettu palvelu, joka on tarkoitettu lokitietojen keräämiseen, yhteenvetoon ja siirtoon. Se on joustava ja vikasietoinen mekanismi, jossa on moninkertaiset vika- ja palautusmenetelmät. Flumen datamalli on yksinkertainen ja sitä käytetään online-analytiikassa. Flume auttaa poimimaan tietoja Twitteristä ja muista sosiaalisista medioista ja tallentamaan ne keskitettyyn tietovarastoon kuten HDFS tai HBase. Flume voi säätää datan tiedonsiirron nopeutta lähteistä kohteisiin ja varmistaa, että data virtaa johdonmukaisesti. Se tukee kontekstiin perustuvia reititystietoja ja koska viestit siirretään useilla kanavilla, sen perille toimitus on taattu. (Vibrant Publishers 2017, 81,96.)

Apache Zookeeper on hajautettu palvelu tiedonsiirron hallintaan liityntäpisteryhmissä. Käsiteltäessä suuria datamääriä se toimii parhaiten hajautetussa järjestelmässä, jossa koko data pilkotaan hallittaviksi lohkoiksi ja jaetaan eri palvelimille käsittelyä varten. Vaikka jokin liityntäpiste pettäisi, koko prosessi ei epäonnistu. Palvelun skaalautuvuuden vuoksi liityntäpisteitä ja laitteita voidaan lisätä häiritsemättä prosessia. Vaikka data on jaettu eri palvelimille, Zookeeper käsittelee sitä yhtenä prosessina. (Vibrant Publishers 2017, 113.)

Apachella on Hadoopiin liittyviä useita muitakin projekteja, kuten Ambari, Avro, Cassandra, Chukwa, Hive, Mahout, Spark, Tez. (The Apache Software Foundation 2018).

12 PALVELUNTARJOAJIEN PILVIPALVELURATKAISUJA

Big dataan liittyviä palveluntarjoajia on useita, joilla on monia erilaisia pilvipalveluissa toimivia ratkaisuja liittyen big datan hyödyntämiseen. Toimijoiden nykyiset big data -ratkaisut perustuvat pitkään kokemukseen ja tietotaitoon big datasta. Tietomäärien lisääntyessä myös ratkaisut kehittyvät.

12.1 IBM

Suuria tietovirtoja pystytään käsittelemään ja analysoimaan IBM InfoSphere Streams:ssä ja IBM PureData for Analytics -ratkaisuilla. Nämä perustuvat hyvin laajaan skaalautuvaisuuteen ja rinnakkaisprosessointiin, joten big data -tietoja pystytään käsittelemään pienellä vasteajalla. IBM InfoSphere Streams:ssä on sisään rakennettuna hallinta- ja kehitysympäristö sekä valmiita algoritmejä ja adaptereita erityyppisille datavirroille. IBM PureData for Analytics on valmisratkaisu, joka sisältää ohjelmistot, tietovarastot ja integroidun serveriympäristön ja on helposti hallittava, kustannuksiltaan tehokas ja käyttöönotettavuudeltaan nopea. (Salo 2013, 61-62.)

IBM on kehittänyt paikallaan pysyvien hajanaisten big data -tietojen visualisointiin, louhimiseen, keräämiseen ja suodattamiseen liiketoimintaympäristöihin soveltuvia Hadoop/MapReduce -ratkaisuja. IBM InfoSphere BigInsights ja IBM Social Media Analytics ovat helppokäyttöisiä, nopeasti käyttöönotettavia selainpohjaisia ratkaisuja, jotka toimivat myös pilvipalveluna. (Salo 2013, 62-63.)

12.2 Amazon Elastic MapReduce (EMR)

EMR on Amazon Web Services:in (AWS) tarjoama pilvipalvelu, joka yksinkertaistaa suurten tietokehysten, kuten Apache Hadoop ja Apache Spark, käsittelyä AWS:ssä, kun käsitellään ja analysoidaan suuria tietomääriä. Käyttämällä näitä kehyksiä ja niihin liittyviä avoimen lähdekoodin projekteja Apache Hive ja Apache Pig, voidaan käsitellä tietoja analyysi- ja liike-elämätarkoituksiin. EMR:llä voidaan muuntaa tietoja ja siirtää

niitä muihin AWS-tietovarastoihin ja tietokantoihin, kuten Amazon Simple Storage Service (Amazon S3) ja Amazon DynamoDB. (Amazon Web Services 2019.)

12.3 Google

BigQuery on valmisratkaisu pilvivarastoihin tallennetun datan visualisointiin. BigQueryä voidaan käyttää selainpohjaisen käyttöliittymän kautta, komentorivityökalulla tai tehdä kyselyjä BigQuery REST API -sovellusrajapinnan kautta käyttämällä esimerkiksi Javaa, .NETiä tai Pythonia. On myös paljon kolmannen osapuolen työkaluja, joita voidaan käyttää BigQueryn kanssa muun muassa visualisointiin. BigQuery tekee nopeita SQL-kyselyitä käyttämällä Googlen infrastruktuurin prosessointitehoa. (Google 2019.)

Google Cloud Dataprep -työkalu on Googlen ja Trifactan yhteistyönä rakentama ja suunnittelema pilviratkaisu, jolla voidaan puhdistaa ja valmistella jäsentelemätöntä dataa. Dataprepillä käsiteltyjä tietoja voidaan käyttää data-analyyseihin ja koneoppimismalleihin. (Alghini 2018.)

Google Cloud Dataprep sulauttaa Trifactan älykkään, käyttäjäystävällisen käyttöliittymän sekä Photon Compute Frameworkin ja integroituu Google Cloud Dataflown kanssa nopeaa ja skaalautuvaa tietojenkäsittelyä varten. Cloud Dataprep antaa Google Cloud -käyttäjille, joilla on asianmukaiset oikeudet käyttää, tutkia ja valmistella monipuolista dataa palveluissa, kuten Cloud Storage ja Big Query, erilaisiin jatkokäyttöihin, mukaan lukien analytiikka. (Wilson 2017.)

Esimerkki Dataprep -työkalun käytöstä .CSV-tiedoston siivouksessa ja uudelleen järjestämisessä on liitteessä.

13 YHTEENVETO JA POHDINTA

Yhä lisääntyvien datamäärien vaihtelevuus, nopeus ja luotettavuus, joita tuottavat muun muassa erilaiset sensorit, älylaitteet, sosiaalinen media ja yritysten tietokannat, joiden datatyypit ovat erilaisia, luovat haasteen datan tallennusratkaisuille ja analysoinnille. Big datasta louhimalla ja sitä analysoimalla saatava tieto edesauttaa pääsemään laajamittaisempaan käsitykseen siitä, miten voidaan kehittää esimerkiksi koneoppimisella luotavaa tekoälyä. Tulevaisuudessa tarvitaan tietokoneita, joissa on enemmän laskentatehoa ja tehokkaampaa datahallintaa big datan hyödyntämiseksi.

Big datan avulla yhteiskuntaa koskevien tilastotietojen muodostaminen ja vertailu on kattavampaa kuin ennen. Yritysmaailmassa big datasta analysoimalla saatavat tiedot auttavat yrityksiä kehittämään strategioitaan ja toimintaedellytyksiään. Big dataan liittyvien teknologioiden kehittyessä, big dataa hyväksikäyttämällä, saadaan yhä enemmän uusia innovaatioita muun muassa julkishallintoon, liike-elämään, lääketieteeseen, teollisuuteen, maatalouteen ja IoT-laitteisiin, joita on käsitelty esimerkkien kautta tässä opinnäytetyössä. Näiden lisäksi big dataa hyödynnetään muun muassa maanpuolustuksessa, koulutuksessa, pankki- ja vakuutusalalla, mediassa ja politiikassa, joita ei käsitelty esimerkein tässä yhteydessä big datan laajuuden vuoksi.

Big dataa voidaan hyödyntää myös seuraavasti aikaisemmin opinnäytetyössä käsiteltyjen esimerkkien lisäksi:

Big dataa voidaan hyödyntää avaruuden tutkimisessa, muun muassa tutkittaessa radioteleskooppien vastaanottamia radiosignaaleja. Signaalit sisältävät paljon dataa, jossa tarvitaan koneoppimisen metodeja ja mahdollisesti uusia algoritmejä signaalien sisältämän informaation ja sen alkuperäisen lähteen selvittämiseksi.

Lääketieteessä big dataa voidaan hyödyntää myös esimerkiksi vertailtaessa maapallon lämpenemisen johdosta jäätiköiden sulamisesta aiheutuvien uusien bakteerilöydöksien yhteneväisyyksiä jo tunnettuihin bakteerikantoihin. Jos löydetään täysin uusia bakteereja, voidaan tutkia miten ja mihin ne vaikuttavat. Todettaessa bakteeri vaaralliseksi, voidaan kehittää keinoja sen leviämisen estämiseksi.

Big dataa voidaan hyödyntää myös esimerkiksi yhdyskuntatieteissä vaikkapa seurantatutkimuksissa. Valitaan suuri joukko ihmisiä, joiden poliittista suuntautumista tutkitaan, muuttuuko se esimerkiksi ikävuosien 18-30 välillä ja mitä vaikutuksia mahdollisiin muutoksiin on sukupuolella, kotikasvatuksella, ympäristöoloilla ja koulutuksella. Otantoja tilastoidaan esimerkiksi kolmen vuoden välein, jolloin saadusta aineistosta voidaan analysoida muutoksiin eniten vaikuttavat tekijät tai niiden yhdistelmät.

Big dataa käyttämällä voitaisiin tutkia eri aineita ja niiden yhdisteitä, jotta markkinoille saataisiin ympäristöystävällinen torjunta-aine, esimerkiksi kaislojen poistoon vesialueilta. Nykyisin markkinoilla olevat torjunta-aineet ovat haitaksi vesistöille ja vesieliöille, joten ainoaksi poistokeinoiksi jää joko ruoppaus tai leikkaus. Big datan avulla voisi mallintaa aineyhdistelmien tehoa, mahdollisia haittoja ja aineen häviämistä luonnosta.

Datan suuren määrän vuoksi se on järkevää tallentaa pilvipalveluun ja käsitellä palveluntarjoajien analytiikkaratkaisuilla. Organisaatioiden, jotka käsittelevät tietoja ”pilvessä”, on hyvä tehdä turvallisuusanalyysi etukäteen ja tämän pohjalta tehdä sopimus ja palvelutasomääritykset pilvipalvelutarjoajan kanssa. GDPR:n voimaantulon jälkeen kuluttajien yksityisyydensuoja on parantunut. Yritysten on ilmoitettava muun muassa evästeiden käytöstä nettisivuillaan ja niiden käyttötarkoitus sekä antaa mahdollisuus evästeiden kieltämiseen esimerkiksi mainosten kohdentamisessa. Yrityksille GDPR aiheuttaa järjestelmämuutoksia, jotka aiheuttavat lisäkustannuksia. EU:n alueella toimivat organisaatiot ja yritykset, jotka käsittelevät henkilötietoja ovat vastuussa tietosuoja-asetuksen noudattamisesta. GDPR onkin useissa maissa lisäasetus maan omiin lakeihin.

Big datan tallentamiseen ja käsittelyyn on useita alustoja. Yksi näistä avoimeen lähdekoodiin perustuva Hadoop, joka toimii myös pilvipalveluissa. Se on suunniteltu suurien datamäärien käsittelyyn ja varastointiin. Apachella on Hadoopiin liittyviä useita projekteja, jotka käyttävät Hadoopin hajautettua tiedostojärjestelmää. Pilvipalveluiden tarjoajia ratkaisuineen on useita, joista tässä opinnäytetyössä on vain IBM:n joitakin ”pilvessä” toimivia ratkaisuja, Amazon Web Servicen EMR ja Googlelta BigQuery ja Cloud Dataprep.

Näkemykseni mukaan, vaikka big datan hyödyntämisen etuna on, että sillä saadaan laajaa ja vertailukelpoista tietoa eri alojen toimintaan ja uusien innovaatioiden kehittämiseen, on siihen liittyvien ongelmien ja väärinkäytön huomioiminen tärkeää. Eräs ongelma big datan hyödyntämisessä on siinä, kuka sitä tuottaa, käyttää ja mihin tarkoitukseen, onko data salaista vai julkista sekä kuka sen omistaa. Trollitehtaiden tuottama data vääristää totuudenmukaisuutta. Terveydenhoidossa salaista tietoa ovat potilastiedot, maanpuolustuksessa siinä käytettävä teknologia ja liike-elämässä yrityssalaisuudet. Tekoälyyn ja robotteihin liittyvänä ongelmana voidaan pitää, että ne suorittavat vain niihin ohjelmoituja tehtäviä eivätkä kykene itsenäisesti eettisiin ratkaisuihin.

Big datan käytön yleistyessä tarvitaan lisää siihen liittyvää koulutusta, jolloin saadaan alalle lisää osaajia, joista on tällä hetkellä vielä pulaa. Uusien yritysten syntyminen on täten myös mahdollista. Big datan käytön lisääntyessä ja laajentuessa, myös määritelmät lisääntyvät, perusasioiden pysyessä ennallaan, esimerkiksi joidenkin alkuperäisten ”V”-termien lisäksi on tullut useita uusia ”V”-termejä.

Big dataan liittyvää kirjallisuutta on paljon muilla kielillä painettuina ja sähköisinä julkaisuina. Suomalaisia ja suomeksi käännettyjä on hankala löytää. Tässä opinnäytetyössä on hyödynnetty kirjoja, blogikirjoituksia ja sähköisiä julkaisuja.

Aihe on mielenkiintoinen, mutta big datan laajuuden vuoksi kaiken kattavaa tietopakettia ei voinut tehdä tässä yhteydessä. Käsittelemättä jäivät muun muassa Tor-verkko, dark web ja deep web, joka on arviolta 400-500 kertaa laajempi kuin perinteisesti selailtava verkko. Opinnäytetyössä käsiteltyjä alueita jouduttiin rajaamaan todella paljon, koska big dataa hyödynnetään lähes kaikilla aloilla. Tulevaisuudessa big datan oletetaan laajenevan entisestään ja sitä hyödyntävien innovaatioiden lisääntyvän eri aloilla.

LÄHTEET

Acieta LLC. 2019. Robotic manufacturing for automobiles. Viitattu 29.7.2019.

https://www.acieta.com/why-robotic-automation/robotic-solutions-industry/automotive-applications/

Alghini, C. 2018. Cleansing Your Big Data Strategy with Cloud Dataprep. Viitattu 14.8.2019.

https://www.coolheadtech.com/blog/cleansing-your-big-data-strategy-with-cloud-dataprep

Amazon Web Services. 2019. What Is Amazon EMR? Viitattu 12.6.2019.

https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-what-is-emr.html

Bayer AG. 2018. Finding a cure for cancer – is big data the solution? Viitattu 28.6.2019.

https://www.canwelivebetter.bayer.com/innovation/finding-cure-cancer-big-data-solution

Carson, A. 2018. Should vendors be able to pass along costs of GDPR compliance? Viitattu 7.8.2019.

https://iapp.org/news/a/should-vendors-be-able-to-pass-along-costs-of-gdpr-compliance/

COMCAST. 2019. Smart Home Technology: The Best Way to Modernize Your Home. Viitattu 15.7.2019.

https://www.xfinity.com/hub/smart-home/smart-home

Concessao, R. 2016. What is big data really? 5.painos. La Vergne, Tennessee: iCS.

Davis, N. 2019. Artificial Intelligence and Big Data: A Powerful Combination for Future Growth. Viitattu 6.7.2019.

https://su.org/blog/artificial-intelligence-and-big-data-a-powerful-combination-for-future-growth/

Gemalto NV. 2019. Secure, sustainable smart cities and the IoT. Viitattu 21.7.2019.

https://www.gemalto.com/iot/inspired/smart-cities

Google. 2019. What is BigQuery? Viitattu 9.8.2019.

https://cloud.google.com/bigquery/what-is-bigquery

Holmes, D. 2017. Big data: A Very Short Introduction. 3.painos. New York: Oxford University Press.

Hurwitz, J., Nugent, A., Halper, F. & Kaufman, M. 2013. Big Data For Dummies. Hoboken, New Jersey: John Wiley & Sons, Inc.

IDC. 2018. The Digitization of the World From Edge to Core. Viitattu 23.2.2019.

https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-da-taage-whitepaper.pdf

Kaushik. 2016. Hadoop key terms, Simplified. Viitattu 23.2.2019.

http://techalpine.com/hadoop-key-terms-simplified/?lang=en

Kolb, J. & Kolb, J. 2013. The Big Data Revolution. La Vergne, Tennessee: Applied Data Labs Inc.

McNeill, C. 2019. Veracity: The Most Important “V” of Big Data. Viitattu 1.7.2019.

https://www.gutcheckit.com/blog/veracity-big-data-v/

META Group. 2011. Application Delivery Strategies. Viitattu 23.2.2019.

https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

NIST. 2011. The NIST Definition of Cloud Computing. Viitattu 23.2.2019.

https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf

NIST. 2015. NIST Big Data Interoperability Framework: Volume 4, Security and Privacy. Viitattu 23.2.2019.

https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-4.pdf

NVIDIA Corporation. 2016. Autonomous Cars. Viitattu 16.7.2019.

https://www.nvidia.com/object/autonomous-cars.html

Patidar, S. 2018. Big Data in Healthcare: Real World Use-Cases. Viitattu 2.8.2019.

https://dzone.com/articles/big-data-in-healthcare-real-world-use-cases

Pickell, D. 2018. What is Big Data? A Complete Guide. Viitattu 6.8.2019.

https://learn.g2.com/big-data

Rai, A. 2019. What is Big Data – Characteristics, Types, Benefits & Examples. Viitattu 3.8.2019.

https://www.upgrad.com/blog/what-is-big-data-types-characteristics-benefits-and-examples/

Salo, I. 2013. Big data tiedon vallankumous. Jyväskylä: Docendo Oy.

Salo, I. 2014. Big data & pilvipalvelut. Jyväskylä: Docendo Oy.

Schmidt, S. 2018. 5 Major Industries That Will Be Changed by Robotics. Viitattu 16.7.2019.

https://blog.marketresearch.com/5-major-industries-that-will-be-changed-by-robotics

Scott, B. 2018. What's in it for Consumers? The Top 5 Privacy Benefits of the GDPR. Viitattu 28.7.2019.

https://blog.centrify.com/consumer-privacy-benefits-gdpr/

Madsen, C., Cormier, E., Von Stecher, J., Liu, K., Voronov, G., Gu, H. & Wiley, E. 2014. Python for Analytics and The Role of R. Viitattu 1.8.2019.

https://www.seagate.com/files/www-content/ti-dm/_shared/images/r-and-python-pv0026-1-1409us.pdf

Techopedia Inc. n.d. Autonomous Car. Viitattu 16.7.2019.

https://www.techopedia.com/definition/30056/autonomous-car

The Apache Software Foundation. 2018. Apache Hadoop. Viitattu 23.2.2019.

http://hadoop.apache.org/

The R Foundation. n.d. What is R? Viitattu 21.7.2019.

https://www.r-project.org/about.html

Tomlinson, Z. 2018. 15 Medical Robots That Are Changing the World. Viitattu 11.7.2019.

https://interestingengineering.com/15-medical-robots-that-are-changing-the-world

University of Cambridge. 2015. Artificially-intelligent Robot Scientist 'Eve' could boost search for new drugs. Viitattu 17.7.2019.

https://www.cam.ac.uk/research/news/artificially-intelligent-robot-scientist-eve-could-boost-search-for-new-drugs

Van Rijmenam, M. n.d. Why The 3V's Are Not Sufficient To Describe Big Data. Viitattu 6.8.2019.

https://datafloq.com/read/3vs-sufficient-describe-big-data/166

Vibrant Publishers. 2017. Hadoop BIG DATA Interview Questions You'll Most Likely Be Asked. 1.painos. Erie, USA: Vibrant Publishers.

Wade, M. 2018. GDPR comes with teeth – here are the winners and losers. Viitattu 7.8.2019.

https://theconversation.com/gdpr-comes-with-teeth-here-are-the-winners-and-losers-96375

Wilson, A. 2017. A New Cloud-Based Data Prep Solution from Google & Trifacta. Viitattu 2.9.2019.

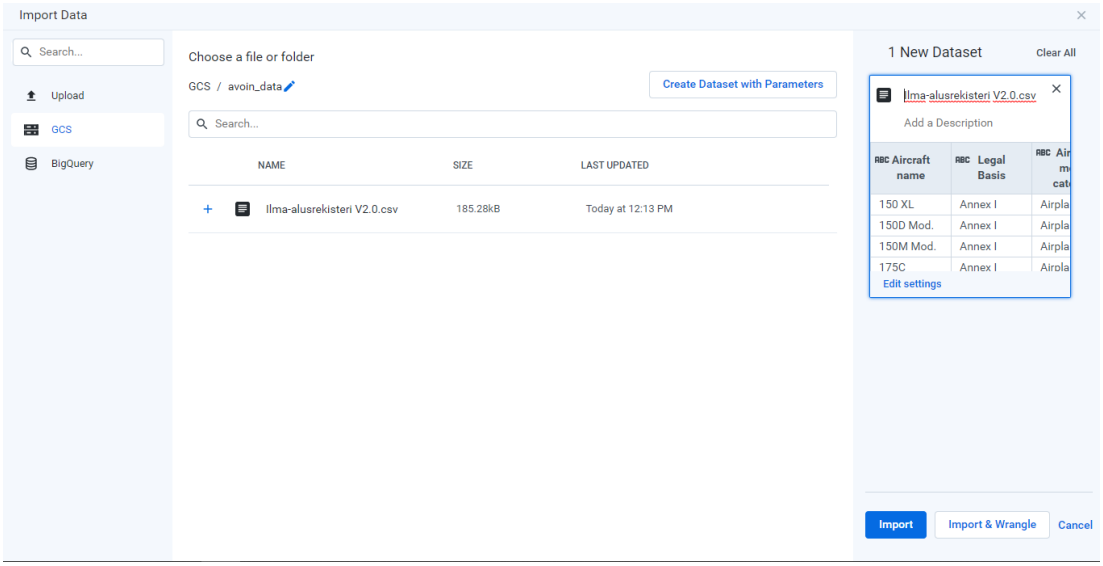
https://www.trifacta.com/blog/data-preparation-solution-google-trifacta/

LIITE

.CSV -TIEDOSTON PUHDISTAMINEN JA JÄRJESTÄMINEN DATAPREP -OHJELMALLA

Käytettävänä datana on Traficomin ilma-alusrekisteri avoin data 2.0. Tiedosto on ladattu ZIP-pakattuna XLXS-tiedostona ja muutettu CSV-tiedostoksi omalla koneella ja ladattu Google Cloud Storageen (GCS). Tarkoituksena on puhdistaa ja järjestää data siten, että saadaan tiedot helikoptereista, niiden ICAO -tyyppitunnus, nimi, valmistusvuosi, moottorien lukumäärä, ensimmäisen ja toisen moottorin valmistaja ja tyyppimerkintä sekä vähimmäismiehistö. Tässä käytetään pientä tiedostoa, mutta myös suurempia voidaan käsitellä samalla tavalla.

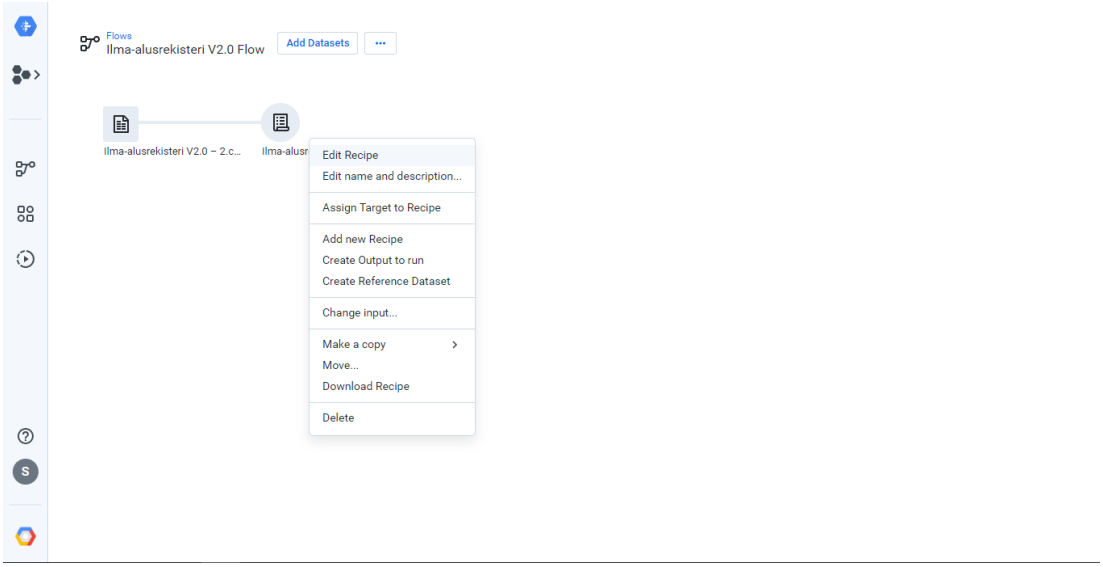
Tiedoston siirto GCS:stä ohjelmaan:



Kuva 4. Tiedoston siirto.

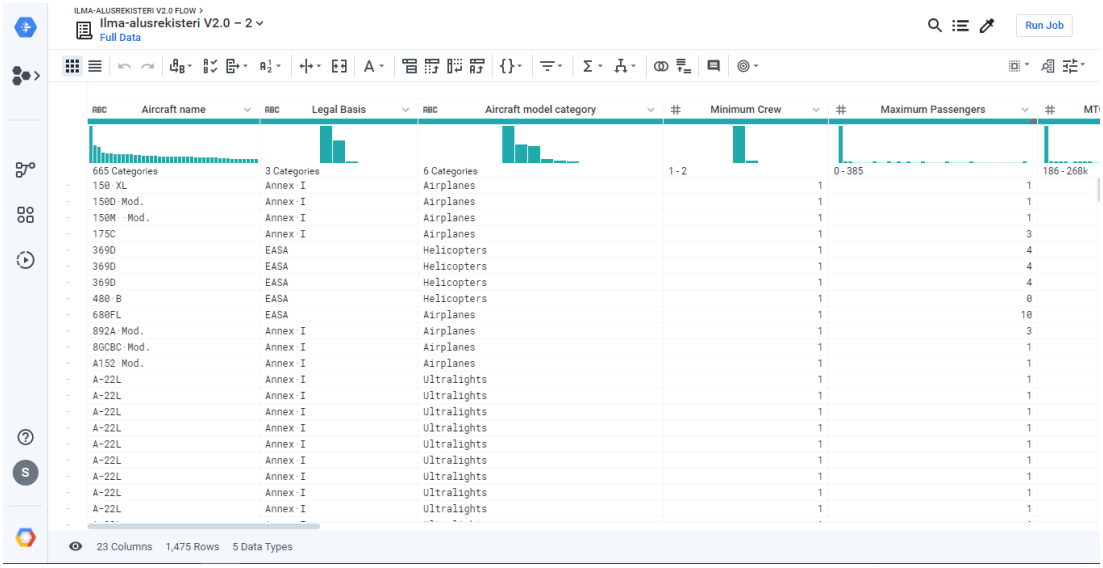
Kuvassa 4 valitaan vasemmalta mistä tiedot halutaan tuoda ohjelmaan. Ohjelma näyttää mitä tiedostoja GCS sisältää. Jos valitaan ”Import” -painike, vain pelkkä tiedosto tuodaan. Kun valitaan ”Import & Wrangle” -painike, ohjelma avaa tiedoston suoraan käsittelyä varten.

Flow ja Recipe:



Kuva 5. Flow ja Recipe.

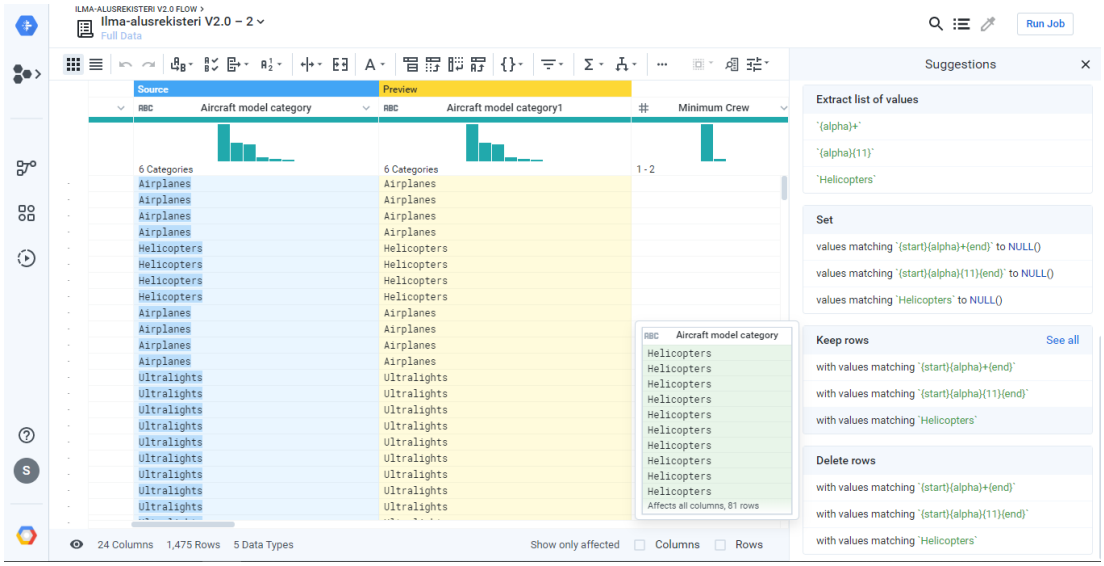
Kuvassa 5 vasemmalla on GCS:stä tuotu tiedosto tietovirtaan, jota ohjelma nimittää ”flowksi”. Jokainen ”flow” voi sisältää yhden tai useampia reseptejä (Recipe), joita käytetään lähdetiedostojen muuntamiseen. Uusia reseptejä luodaan valitsemalla ”Add new Recipe”. Reseptejä käsitellään valitsemalla ”Edit Recipe”, ohjelma avaa tiedoston käsittelyä varten.



Kuva 6. Recipen avaama näkymä.

Kuvassa 6 Recipen avaama näkymä. Vasemmalla ylhäällä ”Full Data” kertoo, että tiedosto on käsittelemätön. Kun tiedostoa on käsitelty, uudessa ”Recipessä” lukee ”Initial Sample”. Vasemmalla alhaalla ovat tiedoston sarakkeiden lukumäärä, rivien määrä ja datatyypit. Pylväsdiagrammit näyttävät prosentuaalisesti, miten sarakkeissa olevat kategoriat jakautuvat.

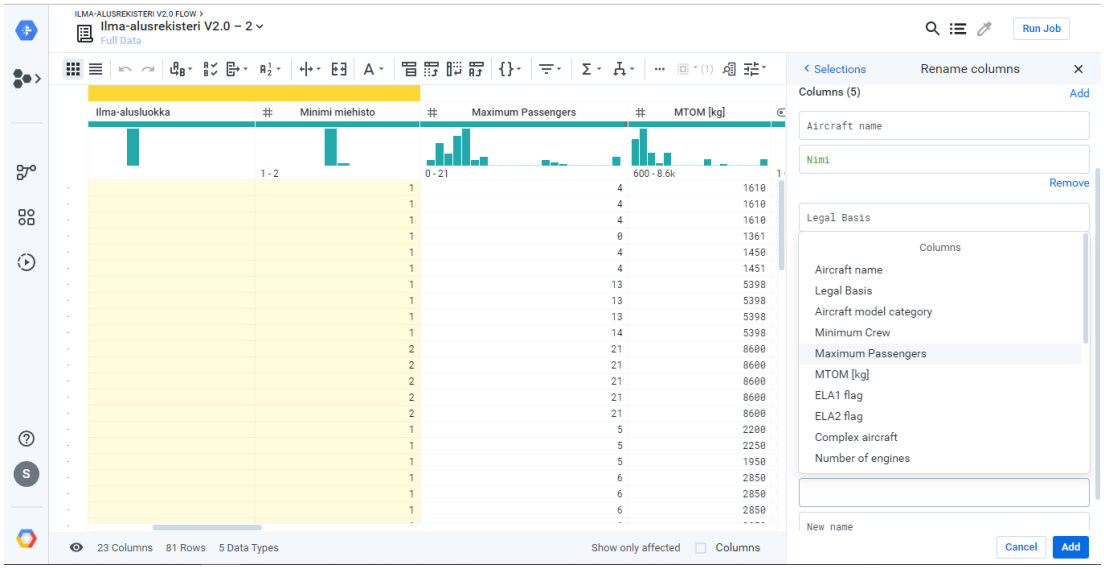
Kohteen valinta sarakkeesta:



Kuva 7. Kohteen valinta sarakkeesta.

Kuvassa 7 on aiemmin “maalattu” Aircraft model category sarakkeesta ”Helicopters” rivi. Ohjelma näyttää ehdotuksia, mitä voi tehdä ”maalatuille” riveille ja mihin sarakkeisiin se vaikuttaa. Kun valitaan ”keep rows with values matching Helicopters”, ohjelma poistaa kaikki muut paitsi helikoptereihin liittyvät rivit.

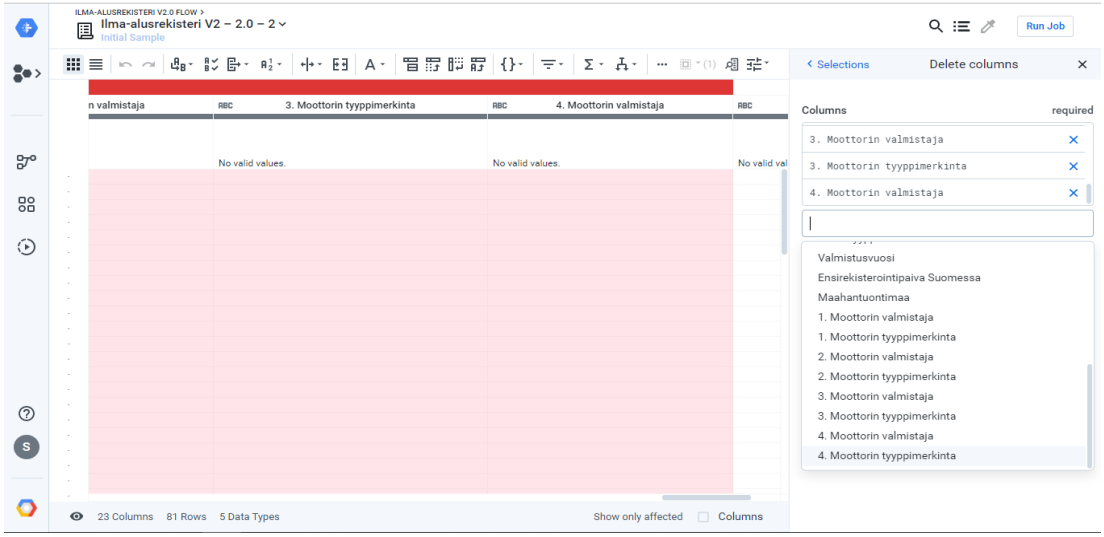
Sarakkeiden uudelleen nimeäminen:



Kuva 8. Sarakkeiden uudelleen nimeäminen.

Kun kuvassa 8 sarakkeiden vieressä olevasta alasvetovalikosta valitaan ”Rename”, ohjelma näyttää nimettävän sarakkeen. Oikeaan reunaan avautuvassa valikossa näkyy sarakkeen vanha nimi, jonka alapuolella olevaan laatikkoon kirjoitetaan uusi nimi. Useampia sarakkeita nimettäessä kerralla painetaan ylempää ”add” -painiketta. Ohjelma näyttää kaikki taulukossa olevat nimet valikkona. Kun halutut sarakkeet ovat nimettyinä uudelleen, painetaan alempaa ”add” -painiketta, jolloin ohjelma lisää uudet nimet taulukkoon.

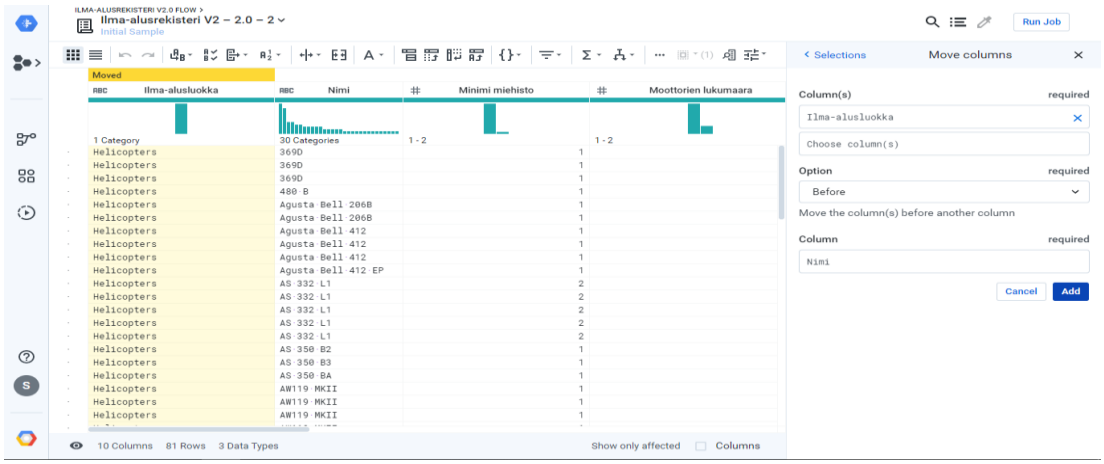
Sarakkeiden poisto:



Kuva 9. Sarakkeiden poisto.

Kuvassa 9 valitaan poistettava sarake “klikkaamalla” sarakenimeä. Oikealle aukeaa ehdotukset ikkuna, josta voidaan valita ”Delete column”, jos halutaan poistaa vain yksi sarake, painetaan ”add”. Useiden sarakkeiden poistossa painetaan ”edit”, avautuu ikkuna, josta voidaan valita useampi poistettava sarake yhdellä kertaa. Sarakkeiden poistaminen tapahtuu samalla periaatteella kuin sarakkeiden nimeäminen.

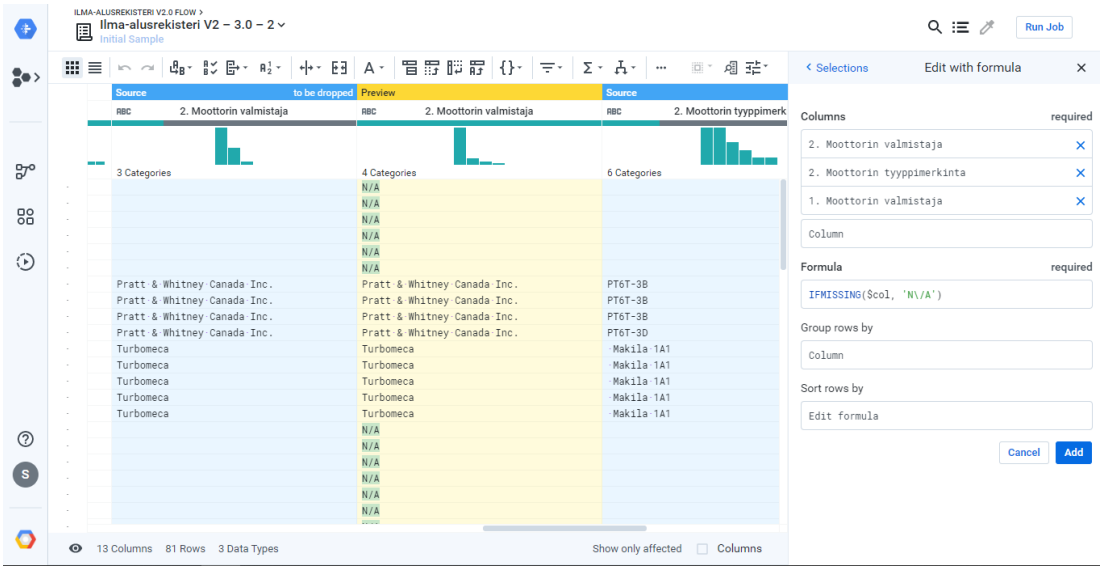
Sarakkeiden siirtäminen:



Kuva 10. Sarakkeiden siirtäminen.

Kuvassa 10 on aikaisemmin valittu ilma-alusluokka sarakkeen alasvetovalikosta ”Move” ja ”To Beginning”. Oikealle aukeaa ”Move columns” -ikkuna, josta voidaan tarkastaa mihin sarake siirretään. Tämä siirto hyväksytään ”add” -painikkeella. Samalla periaatteella kuin sarakkeiden poisto, voidaan useamman sarakkeen paikkaa vaihtaa.

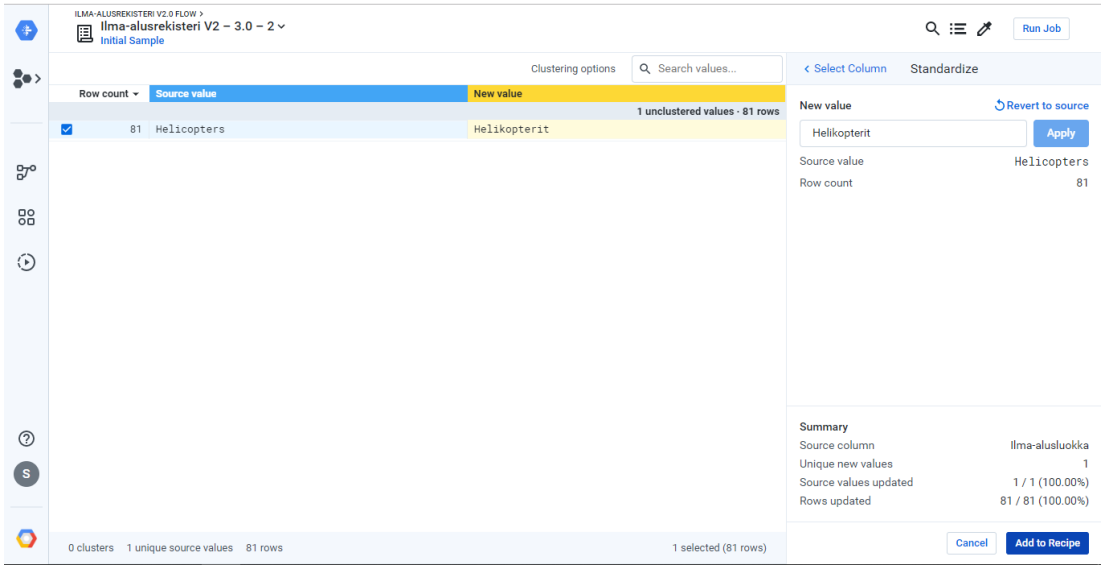
Sarakkeiden tyhjien rivien käsittely:



Kuva 11. Tyhjien rivien käsittely.

Kuvassa 11 sarakenimien alapuolella oleva vaakapalkki näyttää sarakkeiden rivien arvoja. Kuvassa 11 olevassa 2. Moottorin valmistaja sarakkeessa on riveiltä puuttuvia arvoja. Klikkaamalla poikkipalkkia oikealle aukeaa ehdotukset-ikkuna, josta voidaan valita mitä puuttuville arvoille tehdään. Tässä taulukossa laitetaan jokaisen sarakkeen tyhjän rivin arvoksi N/A. Valitaan ”set” kohdasta ”missing values to N/A” ja klikataan ”edit”. Oikealla avautuvassa ”Edit with formula” -ikkunassa voidaan valita sarakkeet, joiden riveille kirjoitetaan N/A ja painetaan ”add”.

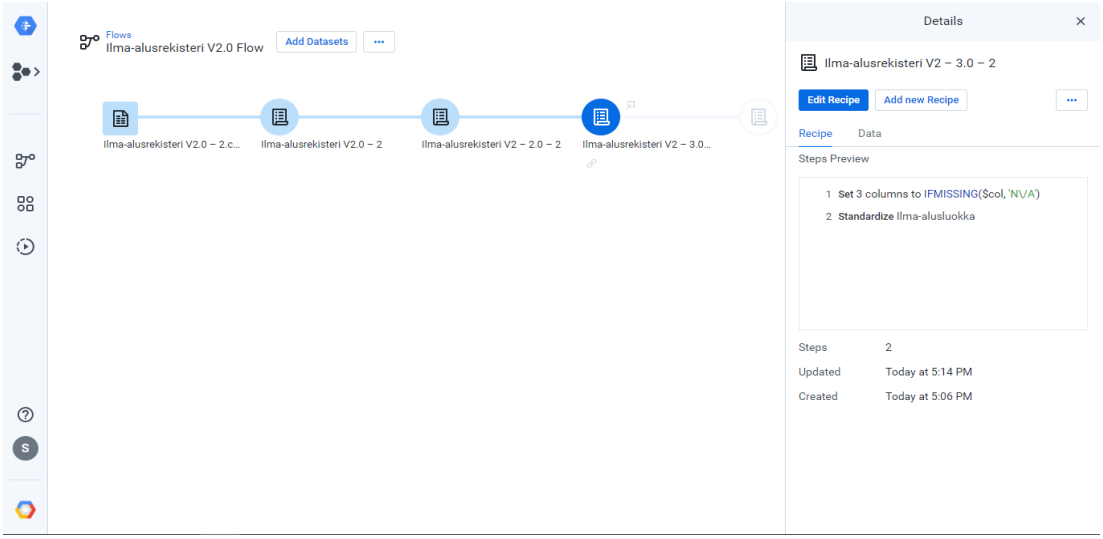
Sarakkeiden rivien nimitietojen muuttaminen:



Kuva 12. Sarakkeiden rivien nimitietojen muuttaminen.

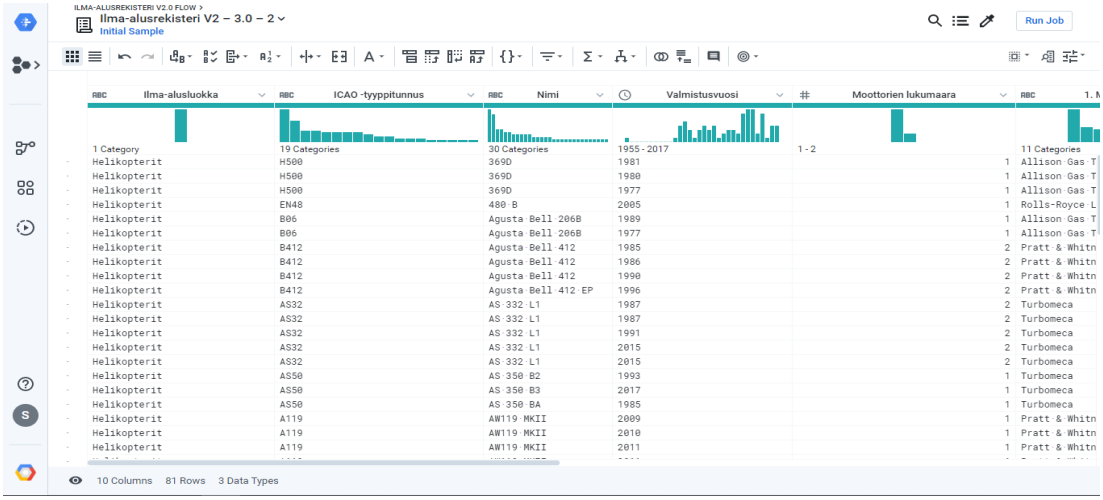
Kuvassa 12 ilma-alussarakkeen alasvetovalikosta on valittu ”Standardize”, ohjelma avaa uuden ikkunan, jossa näkyvät kyseisen sarakkeen rivien nimitiedot ja kuinka monella rivillä ne ovat. Vasemmalta rastitetaan ”Row Count” -kohdasta haluttu muokattava rivi. Oikealla ylhäällä kohtaan ”New value” annetaan riveille uusi tieto ja hyväksytään se ”Apply” -painikkeella, jolloin oikealla alhaalla näkyy yhteenveto tulevista muutoksista, jotka hyväksytään ”Add to Recipe” -painikkeella.

Run job:



Kuva 13. Flow näkymä ennen Run Jobia.

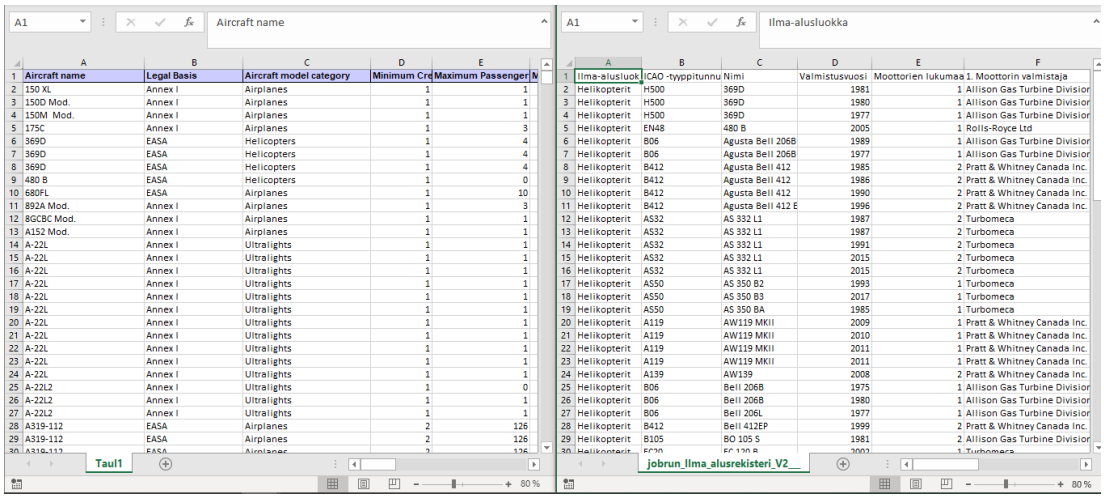
Kuvassa 13 näkymä ”resepteistä”, joista selviää käsittelyvaiheet. Klikkaamalla ”reseptiä”, oikealla näkyy mitä kyseisessä kohdassa on tehty.



Kuva 14. Run Job

Kuvassa 14 on tehty halutut muutokset taulukkoon. Tämän jälkeen painetaan oikeasta yläkulmasta ”Run Job” -painiketta, jolloin ohjelma kysyy minne ja missä formaatissa tiedosto tallennetaan. Ohjelman kirjoittamisen formaatteja ovat Avro, CSV ja JSON.

Lopputulos avattuna Excelissä:



Kuva 15. CSV-tiedosto avattuna Excelissä.

Kuvassa 15 vasemmalla alkuperäinen tiedosto ja oikealla tiedosto muokattuna.