

Sofia Korhonen

Data warehousen käyttöönotto raportoinnin tarpeisiin: case Barona Logistiikkaratkaisut Oy

Tradenomi
Tietojenkäsittely
Syksy 2019



**KAMK • University
of Applied Sciences**

Tiivistelmä

Tekijä: Korhonen Sofia

Työn nimi: Data warehousen rakentaminen raportoinnin tarpeisiin: case Barona Logistiikkaratkaisut Oy

Tutkintonimike: tradenomi (AMK), tietojenkäsittely

Asiasanat: data warehouse, business intelligence, raportointi, AWS, Power BI

Tässä opinnäytetyössä käydään läpi data warehousen rakentamisprojektia Barona Logistiikkaratkaisut Oy:lle. Data warehouse rakennetaan erityisesti palvelemaan operatiivisen raportoinnin tarpeita. Logistiikkaratkaisut teki jo aikaisemmin raportointia Microsoftin Power BI -ohjelmalla, mutta siihen liittyi haasteita esimerkiksi datalähteiden erilaisuuden ja roskaisuuden takia. Data warehousen rakentaminen nopeuttaa myöhempää raportointityötä, kun aikaa ei tarvitse enää käyttää jokaisen datalähteen siivoamiseen.

Projektin pääpalveluiksi valittiin Amazon Web Services ja Microsoft Power BI, joita verrattiin muutamiin muihin pilvipalveluihin sekä business intelligence -työkaluihin. Nämä palvelut valittiin niiden ominaisuuksien, hinnoittelun sekä konsernitason vaikutusten takia. Lisäksi opinnäytetyössä avataan business intelligenceen ja data warehousen käsitteitä, miten ne liittyvät toisiinsa ja miten ne auttavat yritystä tekemään parempaa raportointia ja parempia businesspäätöksiä. Työn lopussa esitellään uutta raportointimallia, joka Logistiikkaratkaisuille tässä projektissa luotiin.

Tämän projektin ensimmäinen tavoite oli luoda asiakaskohtainen tehokkuusraportointi uuden data warehousen päälle. Tähän tavoitteeseen päästiin hyvin ja uusi raportointimalli todettiin onnistuneeksi. Kun kaikkien asiakkaiden data on nyt käsitelty yhtenäiseen muotoon, on sitä helppo lähteä jalostamaan eteenpäin myös uusissa projekteissa.

Abstract

Author: Korhonen Sofia

Title of the Publication: Building a Data Warehouse for Reporting: Case Barona Logistiikkaratkaisut Oy

Degree Title: Bachelor of Business Information Technology

Keywords: business intelligence, data warehouse, reporting, AWS, Power BI

The objective of this Bachelor's thesis is to describe building a data warehouse for Barona Logistiikkaratkaisut Oy. The data warehouse will especially serve the needs of operative reporting. Logistiikkaratkaisut already did some reporting with Microsoft Power BI but there were challenges with it, for example differences and untidiness of data sources. Building a data warehouse will speed up later reporting as time won't be needed to clean each data source separately.

The main services chosen for this project were Amazon Web Services and Microsoft Power BI. They were compared to some other cloud services and business intelligence tools. These services were chosen for their features, pricing and effect on other Barona companies. Additionally, this thesis covers the definitions of business intelligence and data warehousing, how they are related to each other and how they will help companies to build better reporting and make better business decisions. The final part of this thesis describes the new reporting model that was built for Logistiikkaratkaisut in this project.

The first goal of this project was to build performance reporting per customer based on the new data warehouse. This goal was reached, and the new reporting model was deemed successful. As all customer data has now been processed into a unified form, it will be easier to utilize in new projects.

Sisällys

| | | |
|-----|---|----|
| 1 | Johdanto | 1 |
| 2 | Business intelligence..... | 2 |
| 3 | BI-työkalut | 3 |
| 3.1 | Microsoft Power BI..... | 3 |
| 3.2 | AWS QuickSight | 4 |
| 3.3 | Tableau | 5 |
| 3.4 | Vertailun tulos | 5 |
| 4 | Data warehousing..... | 6 |
| 5 | Pilvipalvelut | 8 |
| 5.1 | Amazon Web Services | 8 |
| 5.2 | Microsoft Azure | 9 |
| 5.3 | Google Cloud Platform | 9 |
| 6 | Projektin esittely: Barona Logistiikkaratkaisut Oy..... | 11 |
| 6.1 | Lähtötilanne | 11 |
| 6.2 | Projektin tavoitteet | 12 |
| 7 | Data warehousen rakentaminen..... | 14 |
| 7.1 | AWS-komponentit..... | 14 |
| 7.2 | Tietokantarakenne | 15 |
| 8 | Raporttien luominen | 17 |
| 9 | Yhteenveto | 19 |
| | Lähteet | 20 |

1 Johdanto

Tämä opinnäytetyö käsittelee data warehousen rakentamisprojektia Barona Logistiikkaratkaisut Oy:lle. Data warehouse rakennetaan erityisesti palvelemaan yrityksen operatiivisen raportoinnin tarpeita. Työssä avataan ensin business intelligencen käsitettä sekä sitä, miten data on sille tärkeää. Sen jälkeen avataan data warehousen määritelmää ja miten data warehouse auttaa toteuttamaan hyvää business intelligence strategiaa. Teorian esittelyn jälkeen käydään läpi, miten projekti käytännössä toteutettiin ja avataan päätöksiä, joita sen aikana tehtiin. Lopuksi kerätään yhteen ajatuksia projektin kokonaisuudesta.

Aloitin työharjoittelujakseni Logistiikkaratkaisulla elokuussa 2018 ja päätyötehtäväkseni muotoutui yrityksen tehokkuusraportoinnin uudistaminen. Logistiikkaratkaisut tuottaa muun muassa ulkoistuspalveluita varastoille ja tarkka päivittäisen työn seuranta on yksi kannattavan bisneksen avaintekijöistä. Huomasin nopeasti, että raportoitava data oli hajallaan ja hyvin erimuotoista eri asiakkaiden välillä. Esimieheltäni syntyi ajatus data warehousen rakentamisesta datan yhtenäistämiseksi, joka oli jo sen kokoinen projekti, että siitä riittäisi materiaalia opinnäytetyön tekemiseen.

Data warehouse -projekti sujui hyvin ja se saatiin toteutettua yrityksessä suunnitellulla tavalla. Tavoitteeseen päästiin: käytössä oleva data on nyt standardimuodossa, jonka vuoksi raportointi ja tiedon analysointi on tämän projektin jäljiltä sujuvampaa

2 Business intelligence

Business intelligence, lyhennettynä BI, tarkoittaa yksinkertaistettuna yritysten harjoittamaa tiedon keräämistä ja analysointia parempien liiketoimintapäätösten tekemiseksi. Termiä käytti tietävästi ensimmäisen kerran Richard Miller Devens 1860-luvulla kirjoittamassaan teoksessa ”Cyclopaedia of Commercial and Business Anecdotes”, jossa hän analysoi englantilaisen pankkiirin Sir Henry Furnesen menestystä. Furnese seurasi tiiviisti aikansa uutisia ja poliittista liikehdintää, jotta ehtisi reagoida niihin ensimmäisenä. Myöhemmin hän tosin käytti tietojaan kyseenalaisempiin tarkoituksiin ja tuli tunnetuksi korruptoituneena rahoittajana. [1.]

Business intelligence ei kuitenkaan tullut yleisempään käyttöön ennen 1900-luvun puoliväliä. Saksalaissyntyisen Hans-Peter Luhnin IBM:lle kirjoittamaa artikkelia vuodelta 1958 pidetään alan merkkipaaluna ja se ansaitsi Luhnille myöhemmin tittelin ”business intelligencen isä”. Luhn oli tekstianalyysin erikoisosaaja, jonka ideoita olivat muun muassa tiedon jakaminen säännöllisesti yksittäisten kyselyiden sijaan, erikoisosaajien kouluttaminen vaikeiden kyselyiden tekemiseen sekä samankaltaisten dokumenttien ehdottaminen lukijalle. [2.]

1900-luvun puolella teknologia asetti suurimmat rajat BI:n kehittämiseksi. Tietokoneet saapuivat vasta yritysmaailmaan ja tekniikat datan tallentamista sekä analysointia varten olivat vielä kehitteillä. Uusi teknologia oli kömpelöä ja vaati käyttäjältään paljon osaamista. Nopeasti sitä kuitenkin lähdettiin kehittämään eteenpäin käyttäjäystävällisempään ja virtaviivaisempaan suuntaan.

Tällä hetkellä BI:n keskeisimpänä kehityskohteena on niin sanottu self-service-malli. Tällä tarkoitetaan, että loppukäyttäjälle halutaan antaa mahdollisimman paljon valtaa ja mahdollisuuksia luoda itse tarvitsemiaan analyyseja. Jotta tähän päästään, täytyy taustalla olevan datan olla yksinkertaisessa ja mahdollisimman puhtaassa muodossa sekä käytettävissä olevien työkalujen olla helppokäyttöisiä. Self-service BI nopeuttaa analyysien luontia, kun raportointi ei aina ole IT-osaston takana.

Vaikka maailmassa käsiteltävän datan määrä kasvaa jatkuvasti, mikä lisää erilaisten BI-ratkaisujen tarvetta, Kauppalehden kolumnisti Markus Mertanen povaa jo business intelligencen kuolemaa [3]. Mertasen kolumni on ehkä hieman liioitteleva, koska oikeastaan hän puhuu business intelligence -termin vanhentumisesta ja vaihtumisesta johonkin toiseen. Riippumatta siitä, millä nimellä yritysten datan analysointia tulevaisuudessa kutsutaan, tulee ala varmasti olemaan relevantti vielä seuraavatkin viisikymmentä vuotta.

3 BI-työkalut

Tutustuin tätä projektia varten kahteen itselleni uuteen BI-työkaluun: AWS QuickSight ja Tableau. Niitä kumpaakin verrattiin tilaajayrityksellä jo käytössä olevaan Microsoft Power BI -ohjelmistoon, jota olin itsekin käyttänyt harjoittelujaksoni aikana. Tavoitteena oli selvittää, onko muissa ohjelmissa sellaisia eroja, joiden takia muuttoa vanhasta ohjelmistosta kannattaa edes harkita.

3.1 Microsoft Power BI

Power BI on Microsoftin kehittämä data- ja businessanalytiikkatyökalu. Sen ensimmäinen yleinen versio julkaistiin kesällä 2015 [4]. Power BI:n pääkomponentit ovat käyttäjän tietokoneella toimiva Power BI Desktop sekä selaimella käytettävä Microsoftin pilvessä toimiva Power BI Service. Nämä osat yhdessä pystyvät tarjoamaan vaativampia datankäsittelyominaisuuksia kehittäjille sekä valmiiden raporttien jakamisen käyttäjille. Koska kyseessä on Microsoftin tuote, toimii se vaivattomasti Microsoftin käyttäjähallinnan ja Office-tuotteiden kanssa.

Power BI:lla on kaksi hinnoittelumallia: Pro ja Premium. Pro-mallissa jokainen käyttäjälisenssi maksaa 8,40 euroa kuukaudessa ja ohjelman pilvipalvelu on täysin Microsoftin hallinnoima. Premium-mallin hinnoittelu alkaa 4200 eurosta kuukaudessa. Tässä mallissa organisaatio saa käyttöönsä itselleen varattua tilaa Microsoftin palvelimilta, tarkemman hallinnan omaan tilaansa ja rajattoman määrän käyttäjiä ilman erillisiä lisenssimaksuja. [5.]

Gartner listasi Power BI:n alansa johtajaksi vuoden 2019 Magic Quadrant for Analytics and Business Intelligence Platforms -raportissaan [6]. Raportissa tarkasteltiin eri analytiikkatyökalujen vision kokonaisuutta sekä suoriutumiskykyä, kuten kuvasta yksi nähdään. Microsoft on ollut 12 vuotta johtajana Gartnerin raportissa [7].



Kuva 1. Gartnerin Magic Quadrant -analyysi [6]

3.2 AWS QuickSight

QuickSight on Amazon Web Services -pilvipalvelun tarjoama työkalu nopeiden analyysien ja datavisualisointien rakentamiseen. Se julkaistiin loppuvuodesta 2016 [8]. AWS:n osana QuickSight hyötyy pilvipalvelun resurssien skaalautuvuudesta sekä helposta yhdistämisestä alustan muihin palveluihin kuten tietokantoihin.

QuickSightin hinnoittelussa palvelun käyttäjät jaetaan raporttien luojiin ja lukijoihin. Raporttien luojista maksetaan 18 dollarin kuukausimaksua, mutta lukijoista maksetaan heidän käyttönsä mukaan niin, että jokainen alkava 30 minuutin istunto maksaa 0,30 dollaria [9]. Tämä hinnoittelumalli tulee edulliseksi varsinkin silloin, kun organisaatiossa on paljon käyttäjiä, joilla on tarve lukea raportteja vain muutaman kerran viikossa tai kuukaudessa.

Gartnerin Magic Quadrant raporttiin QuickSight ei ole päässyt mukaan [6]. AWS sisältää tätä nykyä todella monia palveluita, eikä QuickSight ole ollut niistä välttämättä näkyvin tai Amazonin kehityksessä tärkein. Koska AWS oli muutenkin käytössämme, pystyin kokeilemaan QuickSightia käytännössä projektin aikana. Oma kokemukseni oli, että sen ominaisuudet olivat melko rajatut ja se tarjosi lukijalle vähän mahdollisuuksia suodattaa raportteja itse haluamallaan tavalla.

3.3 Tableau

Tableaun ensimmäinen versio julkaistiin jo vuonna 2004, joten sillä on selkeä etumatka kehityksajan suhteen [10]. Gartnerin raportissa Tableau on sijoitettu lähes yhtä korkealle Power BI:n kanssa suorituskkyä vertaillen. Vision kokonaisuudessa Tableau jää kuitenkin jonkin verran jälkeen. Joka tapauksessa Tableau ja Power BI ovat molemmat selkeästi johdossa muihin Gartnerin listaamiin tuotteisiin nähden [6].

Hinnoittelussa Tableau on kallein tässä verratuista tuotteista. Käyttäjälisenssien hinta riippuu hieman siitä, haluaako organisaatio käyttää raporttien julkaisemiseen omaa ympäristöään vai Tableaun hallinnoimaa ympäristöä. Itse hallinnoidussa ympäristössä tarvitaan vähintään yksi 70 dollaria kuukaudessa maksava Creator lisenssi sekä vähintään sata 12 dollaria kuukaudessa maksavaa Viewer lisenssiä. [11.]

Ottaen huomioon lisenssien kustannuksen, tarvittavan infrastruktuurin kustannukset sekä ajan, joka uuden ohjelmiston opetteluun ja käyttöönottoon kuluu, Tableauta ei tässä tilanteessa kannattanut edes lähteä testaamaan käytännössä.

3.4 Vertailun tulos

Tämän opinnäytetyön aikana tehty BI-työkalujen vertailu oli melko suppea, mutta tulini kuitenkin siihen tulokseen, että Power BI:ta ei kannata tässä tilanteessa vaihtaa. Power BI:ssa ei koettu suuria vikoja yrityksen käytössä, eikä vertailluissa ohjelmistoissa ollut suuria etuja siihen nähden. Pitää ottaa myös huomioon, että tämän opinnäytetyön tilaajayritys on osa isompaa konsernia. BI-ohjelmiston vaihto vain yhdessä yrityksessä ei olisi järkevää yleisen hallinnon kannalta ja koko konsernin vaihtaminen uuteen ohjelmistoon olisi todella iso projekti, varsinkin kun siihen ei erityisesti koettu tarvetta.

4 Data warehousing

Järkevien businesspäättösten taustalle tarvitaan dataa ohjaamaan näiden päätösten tekemistä. Yksi vaihtoehto datan säilyttämiseen ja tarjoamiseen organisaation käyttöön on data warehousen rakentaminen. Bill Inmon oli ensimmäisten joukossa määrittelemässä data warehousen käsitettä 1970-luvun loppupuolella. Jos Hans-Peter Luhnia pidetään business intelligencen isänä, niin Inmon ansaitsi itselleen nimen data warehousen isä. Laajempaan käyttöön data warehousing tuli 80-luvun puolella erityisesti IBM:n kehitystyön ansiosta. [12.]

Microsoftin järjestelmäarkkitehti James Serra kirjoittaa blogissaan, että data warehousen tehtävä on yksinkertaistettuna ”toimia pysyvänä säilytyspaikkana datalle, jota voidaan käyttää tukena raportointiin, analysointiin ja muihin BI-tehtäviin” [13]. Data warehouse koostuu yleensä jonkinlaisesta tietokannasta sekä niin kutsutusta ”Extract, Transform and Load (ETL)” -ohjelmasta. ETL-ohjelman tehtävänä on hakea dataa sen lähdetiedostosta tai -tietokannasta ja siirtää se data warehousen tietokantaan. Ohjelma voi myös esimerkiksi muokata dataa tai siivota sitä, jotta uudessa tietokannassa data on mahdollisimman yhtenäisessä ja helposti käsiteltävässä muodossa.

Vaikka data warehousen käyttäminen tarkoittaa, että data kopioidaan käytännössä kahteen eri paikkaan, on tällaisella järjestelmällä paljon etuja. ETL-ohjelma mahdollistaa datan käsittelyn ennen käyttäjää, jolloin sitä voidaan siivota ja muuttaa käyttäjälle ymmärrettävämmäksi. Helposti ymmärrettävä data tarkoittaa, että useammat käyttäjät voivat tuottaa itse tarvitsemiaan raportteja, mikä nopeuttaa työskentelyä ja vähentää IT-osaston kuormitusta. Data warehouse voidaan myös optimoida paremmin tukemaan käyttäjien kyselyitä kuin monet yksittäiset datalähteet. Data warehouse voi lisäksi parantaa tietoturvaa, koska käyttöäioikeuksia vain yhteen paikkaan on helppo hallita, sekä helpottaa erilaisten lakien ja säädösten noudattamista, koska kaikki data kerätään yhteen paikkaan, jossa sitä voidaan säilyttää niin kauan kuin on tarpeellista. [13.]

Haastattelin tässä projektissa työskennellyttä konsultti Ville Nahkuria hänen kokemuksistaan data warehousingin parissa. Nahkuri on työskennellyt it-alalla parikymmentä vuotta ja perusti oman yrityksensä Isodata Oy:n vuonna 2017. Nahkuri toteaa, että vaikka data warehouse ei sinänsä ole käsitteenä tai tekniikkana uusi, sen yleistymistä on aikaisemmin hidastanut erityisesti käyttöönoton kalleus. Esimerkiksi Oraclen vuoden 2007 hintalistassa Data Warehousing Express Serverin lähtöhinnaksi ilmoitetaan 40000 dollaria per prosessori [14]. Kun tähän lisää tarvittavan infra-

struktuurin kustannuksen, ei data warehousea kannata lähteä rakentamaan kovin pienen projektin takia. Nykyisessä pilvipalvelumaailmassa kynnys lähteä kokeilemaan on madaltunut ja yritykset voivat lähteä rohkeammin kokeilemaan erilaisia ratkaisuja.

5 Pilvipalvelut

Pilvipalvelumarkkinaa dominoi tällä hetkellä kaksi yritystä: Amazon sekä Microsoft. Näiden jälkeen isoimmat toimijat ovat Alibaba, Google ja IBM. Yhteensä nämä viisi yritystä kattavat yli 75 prosenttia markkinasta, josta lähes 50 prosenttia on Amazonin hallussa, kuten taulukossa 1 esitetään. Kiinalainen Alibaba ei ollut houkutteleva vaihtoehto palveluntarjoajaksi, joten tässä opinäytetyössä tutustutaan Amazonin, Microsoftin ja Googlen pilvipalvelualustoihin. [15.]

| Company | 2018 Revenue | 2018 Market Share (%) | 2017 Revenue | 2017 Market Share (%) | 2018-2017 Growth (%) |
|----------------|-------------------------|--------------------------------------|-------------------------|--------------------------------------|---------------------------------|
| Amazon | 15,495 | 47.8 | 12,221 | 49.4 | 26.8 |
| Microsoft | 5,038 | 15.5 | 3,130 | 12.7 | 60.9 |
| Alibaba | 2,499 | 7.7 | 1,298 | 5.3 | 92.6 |
| Google | 1,314 | 4.0 | 820 | 3.3 | 60.2 |
| IBM | 577 | 1.8 | 463 | 1.9 | 24.7 |
| Others | 7,519 | 23.2 | 6,768 | 27.4 | 11.1 |
| Total | 32,441 | 100.0 | 24,699 | 100.0 | 31.3 |

Source: Gartner (July 2019)

Taulukko 1. Pilvipalveluiden markkinaosuudet ja kasvu 2017-2018 [15]

5.1 Amazon Web Services

Amazon Web Services (AWS) on verkkokauppajätti Amazonin vuonna 2006 lanseeraama pilvipalvelu. Yli 15 miljardin dollarin liikevaihdolla vuonna 2018 AWS on tällä hetkellä selkeästi isoin palveluntarjoaja markkinoilla. Vaikka AWS tuottaa vain reilut 10 prosenttia Amazonin liikevaihdosta, on sen osuus Amazonin liikevoitosta hieman yli puolet [16].

Gartnerin vuoden 2019 Magic Quadrant for Cloud Infrastructure as a Service, Worldwide -raportissa AWS:n vahvuuksiksi kerrotaan muun muassa sen asiakkaiden laajuus ja monimuotoisuus, jolloin teknologiset edelläkävijät auttavat uusien ominaisuuksien pilotoinnissa kaikille asiakkaille, sekä yritysten luottamus AWS:sään, joka näkyy suurissa vuosittaisissa sijoituksissa sekä kriittisten

toimintojen siirrossa AWS:n pilveen. AWS:n mahdolliseksi vaaraksi Gartner kertoo, että tarve olla markkinan ensimmäisenä julkaisemassa uusia palveluita tarkoittaa välillä keskeneräisten palvelujen julkaisua, joita joudutaan korjaamaan vielä pitkään jälkikäteen. AWS:n hinnat eivät myöskään ole laskeneet vuoden 2014 jälkeen, vaikka komponenttien hinnat markkinoilla ovat laskeneet. [17.]

5.2 Microsoft Azure

Microsoftin pilvipalvelualusta Azure aloitti markkinoilla 2013 ja on siitä kasvanut alan toiseksi suurimmaksi toimijaksi yli viiden miljardin dollarin liikevaihdolla vuonna 2018 [15]. Azuren kasvu on hidastunut jonkin verran vuoden 2019 aikana tippuen ensimmäisen kvartaalin 76 prosentista loppuvuotta kohden 64 prosenttiin. Kasvun hidastuminen kertonee kuitenkin enemmän palvelun kasvaneesta koosta eikä Azuren tulevaisuudesta tarvitse olla tässä vaiheessa huolestunut [18].

Gartner toteaa Azuren suurimmaksi eduksi Microsoftin valmiiksi laajan kantaman it-markkinoilla, jolloin sen on helppo lähteä myymään Azurea muiden palvelujen joukossa. Esimerkiksi .NET-pohjaisten sovellusten kehittäminen alusta loppuun kokonaan Microsoftin tuotteiden varassa on tehty todella helpoksi. Azure on myös panostanut avoimen lähdekoodin sovellusten lisäämiseen palveluun sekä yhteistyöhön muiden teknologiayritysten kuten VMwaren ja Red Hatin kanssa. Gartner varoittaa kuitenkin Azuren palvelun luotettavuudesta. Syyskuun 2018 jälkeen Azuressa on ollut monia merkittäviä katkoja, joiden haittojen minimoimiseksi asiakkailta on ollut vain vähän työkaluja. Yritykset ovat olleet myös pettyneitä Microsoftin asiakastuen laatuun ja hintaan. [17.]

5.3 Google Cloud Platform

Hakukonejätti Googlen IaaS-palvelu Google Cloud Platform (GCP) julkaistiin yleisille markkinoille vuoden 2013 lopulla. Vuoden 2018 loppuun mennessä se oli ehtinyt kerryttää itselleen 4 prosentin liikevaihto-osuuden koko pilvipalvelumarkkinasta kasvattaen liikevaihtoaan yli 60 prosenttia vuodesta 2017 [15].

GCP:n vahvuudet ovat Gartnerin mukaan erityisesti big data- ja analytiikkasovelluksissa. Sen asiakkaat voivat hyötyä Googlen keksinnöistä esimerkiksi koneoppimisen parissa. Vaikka aluksi GCP oli suuntautunut pilvinatiivien start up -yritysten palveluun, on se sittemmin laajentunut myös

isojen yritysten palvelemiseen. Yritysassiakkaiden palvelu on Gartnerin mukaan silti edelleen yksi GCP:n heikkouksista, kuten myös vähäinen kokeneiden ammattilaisten määrä kilpailijoihin nähden. Tämä voi käännä osan potentiaalisista asiakkaista pois GCP:n parista, koska tarvittavaa osaamista palvelun vaihtoon ei löydy. [17.]

Tässä projektissa pilvipalvelun tarjoajaksi valikoitui Amazon Web Services. Pelkästään data warehousea varten mikä tahansa näistä palveluista olisi ollut täysin kelvollinen. Opinnäytetyön tilaajayritys on kuitenkin osa isompaa konsernia, jonka eri yrityksillä on erilaiset tarpeet it-palveluille. AWS:n pitäisi pystyä tarjoamaan kaikille näille yritykselle niiden tarvitsemat palvelut sekä nyt että tulevaisuudessa. Markkinajohtajana AWS:n tuntevia työntekijöitä on helppo löytää, halutaanpa heitä sitten rekrytoida omaan yritykseen tai käyttää konsultteina.

6 Projektin esittely: Barona Logistiikkaratkaisut Oy

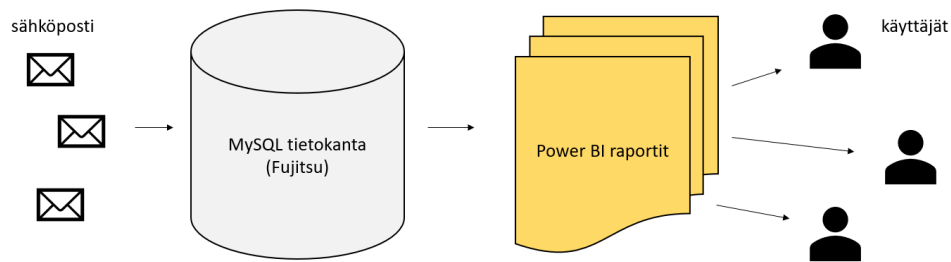
Barona on vuonna 1999 perustettu henkilöstövuokrauspalveluja tuottava suomalainen yritys. Nykyisin Barona on osa yli kahdenkymmenen eri aloilla toimivan yrityksen Bravedo-yritysyhteisöä [19].

Barona Logistiikkaratkaisut Oy, Barona Logistiikka Oy sekä Avain Logistiikka Oy muodostavat Baronan logistiikka-alan palveluita tuottavan osan. Barona Logistiikan tuottamia palveluita ovat muun muassa henkilöstövuokraus, logistiikan ulkoistukset sekä logistiikan asiantuntijapalvelut [20]. Tämä projekti tehtiin Barona Logistiikkaratkaisut Oy:lle. Työntekijöiden määrä Ratkaisuiden asiakaskohteissa vaihtelee noin parin kymmenen ja parin sadan välillä. Kannattavan liiketoiminnan tekemiseksi on kohteiden prosesseja ja työtehokkuutta pystyttävä seuraamaan jollain tavalla. Päätöksiä pitää pystyä tekemään oikean tiedon eikä mahdollisimman lähellä olevien arvioiden pohjalta.

Tässä projektissa työskenteli minun lisäksi yksi ulkopuolinen konsultti sekä muutamia Baronan toimihenkilöitä sekä Logistiikan että IT:n puolelta. Projekti jakautui käytännössä kahteen osaan. Ensimmäinen osa oli konsultin kanssa tapahtuva data warehousen rakentamisvaihe. Toinen osa oli raporttien luominen uuden tietokannan päälle. Ensimmäiseen osaan varattiin konsultille työsken- kentelyaikaa noin kuukausi ja valmistumisperusteeksi määriteltiin se, että data warehouse olisi vaiheessa, jossa Baronan omien työntekijöiden taidot riittäisivät projektin tuotantoon viemiseen. Toiselle osalle ei asetettu tarkkaa aikamäärettä, mutta ainakin osan uusista raporteista piti olla käytössä vuoden 2018 lopussa.

6.1 Lähtötilanne

Ennen tätä projektia Ratkaisuiden datan keruu, prosessointi ja raportointi oli järjestetty kuvan 2 esittämällä tavalla.



Kuva 2. Logistiikkaratkaisuiden raportointi ennen projektia

Edellisen päivän suoritteet sisältävät tiedostot tulivat asiakkailta sähköpostiin, josta Apache Camel -sovellus poimi ja siirsi ne Fujitsun konesalissa sijaitsevaan MySQL-tietokantaan. Power BI:lla otettiin yhteys tähän tietokantaan, haettiin tiedot, käsiteltiin tiedostot järkevään muotoon ja tuotettiin niistä raportteja.

Ongelma oli, että jokainen tiedosto oli täysin erilainen, mutta loppujen lopuksi niistä kaikista haettiin melko samanlaisia tietoja irti. Perusmittarit ovat kuka teki mitä, kuinka paljon ja kuinka kauan siihen kului aikaa. Power BI:n datankäsittelyominaisuudet ovat melko hyvät, mutta ne eivät pysty kaikkeen. Jos kaikkien asiakkuuksien tiedot tuodaan yhteen Power BI -tiedostoon, alkaa tiedosto olla hidas käsitellä. Jos jokaisen asiakkuuden hajauttaa omaan tiedostoonsa, on niiden hallinta paljon työläämpää.

Verkotus Fujitsun pilven kanssa aiheutti myös tiettyjä haasteita. Joko raportit piti julkaista etäyhteyden läpi Fujitsun pilvessä pyörivältä Windows Serveriltä, tai lokaalin SSH-tunnelin piti olla auki jokaisen datapäivityksen yhteydessä. Tämä tarkoitti manuaalisia toimia joka päivä uusimman datan saamiseksi raporteille.

6.2 Projektin tavoitteet

Projektin päätavoite oli tehdä Logistiikkaratkaisuiden raportoinnista mahdollisimman helppoa. Isoin yksittäinen asia tämän toteuttamiseen oli tiedon yhtenäistäminen mahdollisimman aikaisessa vaiheessa. Näin varsinaisia raportteja tehdessä ei tarvitse enää keskittyä niin sanottuun data cleaningiin eli tiedon saattamiseen käytettävään muotoon.

Projektin ohessa Ratkaisuiden it-infrastruktuuri siirrettiin Fujitsun konesalista Amazonin pilvipalvelu Amazon Web Serviceen. Tämä oli osa konsernitasoista it-palvelujen yhdistämistä ja uudistamista. AWS antaa Baronalle paremman kontrollin eri it-palveluihin sekä tarjoaa paljon uusia kehitysmahdollisuuksia.

7 Data warehousen rakentaminen

Tätä projektia lähdettiin rakentamaan Amazon Web Service -pilvipalveluun. Konsultti Ville Nahkuri totesi, että AWS on tämän hetkinen pilvipalveluiden markkinajohtaja ja sieltä löytyy varmasti kaikki meidän mahdollisesti tarvitsemamme ominaisuudet sekä nyt että tulevaisuudessa. Myös tiedon jakamisen kannalta yhteen palveluun keskittyminen on hyödyllistä. Nyt konsernin it-yksiköstä löytyy AWS:n paremmin tuntevia henkilöitä, joilta Logistiikkaratkaisut voi tarpeen tullen kysyä neuvoa ongelmiinsa. Jos Logistiikkaratkaisut olisi siirtänyt omat it-palvelunsa jollekin eri alustalle, ei apua olisi välttämättä löytynyt oman yrityksen sisältä.

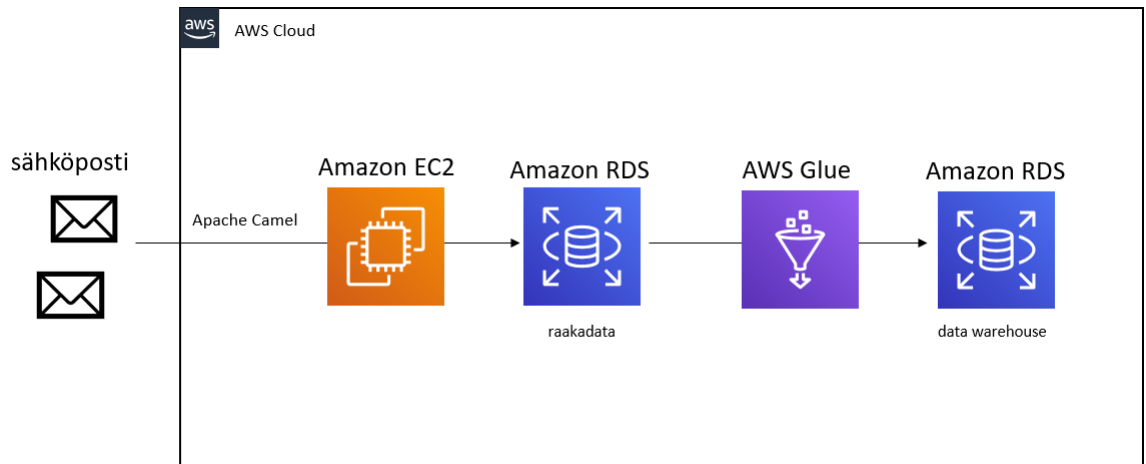
7.1 AWS-komponentit

Kolme AWS-pääkomponenttia, joita päädyimme käyttämään data warehousen rakentamisessa, olivat EC2-instanssi, RDS-tietokanta sekä Glue ETL-putki. EC2 eli Elastic Compute Cloud on AWS:n vuonna 2006 julkaistu virtuaalikonepalvelu [21]. Se on yksi AWS:n ensimmäisistä osista. Tässä projektissa EC2-instanssi pyöritti Apache Camel -sovellusta, joka käy hakemassa datan esimerkiksi asiakkaan rajapinnasta tai sähköpostin liitteestä ja siirtää sen raakamuodossaan ensimmäiseen tietokantaan.

RDS eli Relational Database Service on AWS:n tietokantapalvelu. Vuonna 2009 julkaistu palvelu on erikoistunut relaatiotietokantojen virtualisointiin [22]. RDS-instansseihin on tarjolla useita eri tietokantaohjelmistoja. Tässä projektissa päädyimme käyttämään MySQL-tietokantaa sen helpon hallinnan ja yleisyyden takia, mikä tarkoittaa, että apua ja ohjeita olisi helposti tarjolla ongelmatilanteissa. Samaan tietokantaan voi tallettaa sekä alkuperäisen raakadatan että käsitellyn warehouse datan. Pilvipalvelun tarjoamat resurssit tarkoittavat sitä, että tietokannan kokoa voi tulevaisuudessa kasvattaa kätevästi, jos kasvava datan määrä sitä vaatii.

Glue on AWS:n ETL-palvelu, joka julkaistiin vuonna 2017 [23]. Gluehun luodaan joko Pythonilla tai Scalalla kirjoitettuja ETL-tehtäviä. Palvelun pyöriminen pilvessä tarkoittaa, että käyttäjän ei tarvitse erikseen huolehtia resurssien jakamisesta eri tehtäville. Käyttäjä myös maksaa käyttämistään resursseista vain silloin, kun Glue on aktiivisessa käytössä. Tässä projektissa päädyimme kirjoittamaan ETL-tehtävät Pythonilla, koska tiimistämme löytyi jo valmiiksi jonkin verran Python-

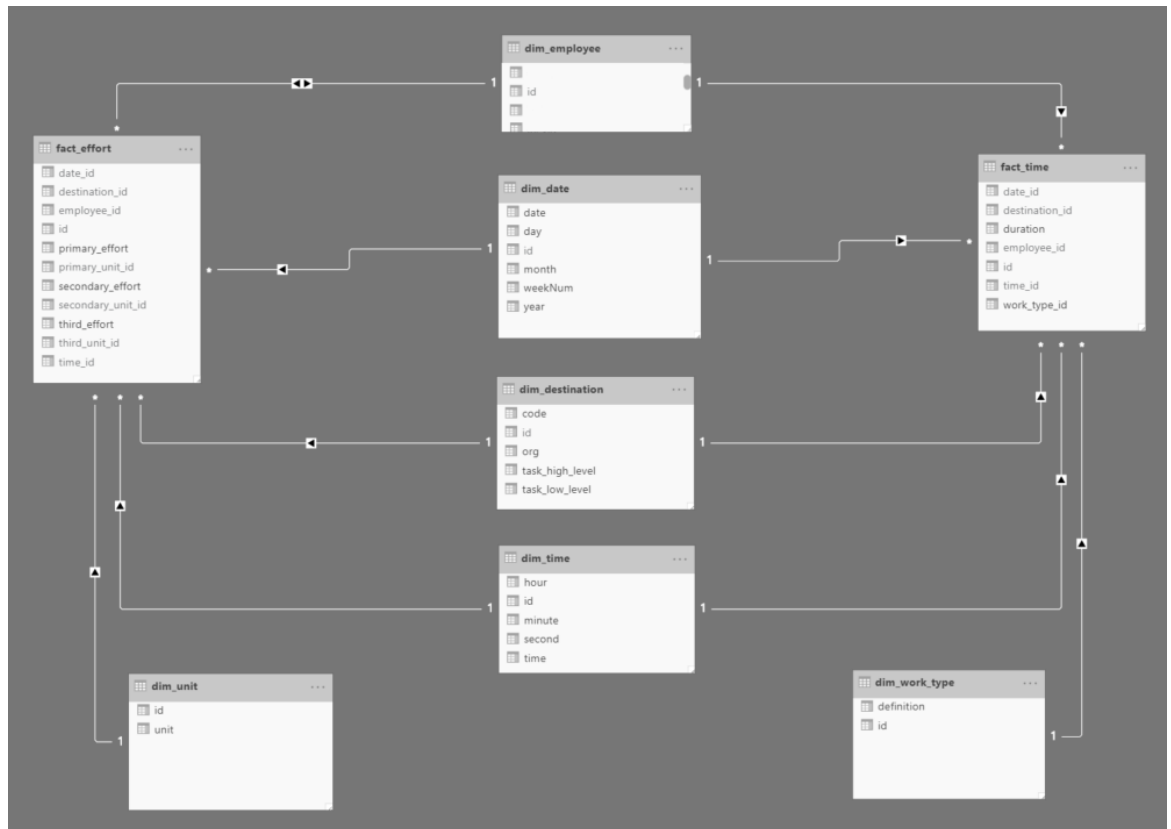
osaamista. Python on muutenkin melko helppo ohjelmointikieli oppia ja sille löytyy runsaasti ohjeita sekä opetusmateriaaleja internetistä. Kuvassa 3 esitetään, miltä projektin AWS-arkkitehtuuri lopulta näytti.



Kuva 3. Projektissa käytetyt AWS-palvelut

7.2 Tietokantarakenne

Data warehousen rakentamista varten piti määrittää tietokannan taulujen rakenne ja niiden suhde toisiinsa. Koska data tulee pääasiassa raportoinnin tarpeisiin, oli tähtimalli tässä projektissa sopiva valinta. Datasta oli selkeästi määriteltävissä kaksi faktaa: aika ja suoritteet. Näihin liittyviksi dimensioiksi määriteltiin päivämäärä, kellonaika, työntekijä ja työtehtävä. Lisäksi pelkästään aikatietoon liittyy palkkalaji-dimensio ja pelkästään suoritetietoon liittyy suoritteiden yksikkö -dimensio. Kuvassa 4 esitellään data warehousen tietokantamalli.



Kuva 4. Data warehousen tietokantamalli. Kuvakaappaus Power BI:sta.

8 Raporttien luominen

Tietokantamallin muututtua kokonaan data warehousen myötä piti raportointi rakentaa uusiksi Power BI:n puolella. Tämä tarjosi mahdollisuuden miettiä raporttien hierarkiaa uudestaan ja virtaviivaistaa raportointia ylläpitotyön minimoimiseksi. Power BI:n pilvipalvelussa raporttien julkaiseminen ja jakaminen toimii Microsoftin työryhmien päällä. Samat ryhmät, jotka organisaatiolle on luotu esimerkiksi SharePointiin, näkyvät myös Power BI:ssa. Ryhmän sisällä kaikki sen jäsenet näkevät kyseiseen ryhmään julkaistut raportit. Lisäksi raportteja voi jakaa yksitellen työtilan ulkopuolisille käyttäjille.

Julkaistaessa Power BI raportti jakaantuu kahteen osaan: visuaaliseen raporttiin ja taustalla olevaan datasettiin. Raportti on se osa, joka sisältää kaikki kaaviot ja visualisoinnit, jotka käyttäjä näkee. Datasetti on raportin taustalla pyörivä osuus, josta raportin sisältämä tieto käydään hake-massa. Sen saa konfiguroitua yhdistymään alkuperäiseen datalähteeseen, josta data voidaan päivittää halutulla aikasyklillä.

Power BI pystyy ottamaan yhteyden moniin erilaisiin datalähteisiin kuten eri tietokantoihin, pilvipalveluihin tai rajapintoihin [24]. Kun datasetti on julkaistu Power BI:n pilveen, voi tätä olemassa olevaa datasettiä käyttää myös uuden raportin datalähteenä. Tämän etu on se, että tarvittavat taulut ja niiden väliset suhteet voidaan tehdä vain kerran yhteen paikkaan. Jos dataa tarvitsee muuttaa, muutos ajautuu yhdestä datasetistä automaattisesti kaikkiin siihen liitettyihin raportteihin.

Logistiikkaratkaisujen kanssa määrittelimme, että jokaisella asiakkaalla olisi hyvä olla oma raporttinsa. Näin raporttia on helppo jakaa henkilöille, jotka työskentelevät kyseisen asiakkaan parissa, mutta joiden ei kuulu nähdä muiden asiakkuuksien raportointia. Jos jokaiselle asiakkaalle luotaisiin oma datasetti, meillä olisi nopeasti yli kymmenen eri datasettiä, joita kaikkia pitäisi hallita ja päivittää erikseen. Yhtä keskitettyä datasettiä käyttämällä taustalla olevan hallinnan määrä vähenee.

Tietoturvan lisäämiseksi otimme käyttöön Power BI:n Row-Level Security (RLS) -ominaisuuden. Power BI -datasettiin luodaan taustalle ”turvallisuusryhmiä”, joihin määritellään, mitä dataa kyseisellä ryhmällä on oikeus nähdä. Datasetin julkaisun jälkeen ryhmään lisätään käyttäjät Power BI:n selaintyökalussa. Turvallisuusryhmiä voi luoda useita ja yksi käyttäjä voi olla jäsen monessa ryhmässä. Power BI osaa myös tulkita eri taulujen välisiä suhteita RLS:n kanssa. Tämän projektin

tapauksessa ryhmät luotiin asiakkaiden perusteella. Power BI osaa rajoittaa automaattisesti sekä aika- että suoritettaa asiakkaan perusteella, kun datasetin suhteet on määritelty oikein. Käyttäjryhmiä luodessa voi olla hyvä luoda yksi master-ryhmä käyttäjille, joilla on oikeus nähdä kaikki data. Tähän käyttäjäryhmään ei tarvitse määritellä minkäänlaisia asetuksia, mutta on helpompaa lisätä käyttäjä yhteen master-ryhmään kuin moneen yksittäiseen ryhmään. Jos käyttäjä ei ole minään ryhmän jäsen, hänellä ei ole näkyvyyttä mihinkään dataan.

Tilaajayrityksen kanssa määrittelimme tämän projektin valmiuskriteeriksi sen, että uuden data warehousen pohjalta saadaan luotua ensimmäinen raportti, jonka pohjalta asiakaskohdetta voidaan johtaa. Tähän tavoitteeseen päästiinkin kohtuullisen nopeasti ja helposti. Yhteensä data warehousen rakentamiseen meni noin puolitoista kuukautta aikaa.

Raportoinnin kehittäminen tietysti jatkuu tämän projektin jälkeen. Eri asiakkuuksien raporttien määrittely vie oman aikansa, jonka jälkeen on tehtävä datan oikeellisuuden validointi. Joskus ETL-ohjelmassa on saattanut tulla virhe kirjoitettuun koodiin, jonka takia data käsitellään väärin. On kuitenkin myös tilanteita, joissa asiakkaalta saatava raakadata ei ole sisältänyt kaikkia raportointiin tarvittavia tietoja. Näiden erojen ja virheiden löytäminen vaatii usein sekä käsittelyn että raakadatan läpikäyntiä, koska raakadataa ei aina siirretä rivi riviltä data warehouseen.

9 Yhteenveto

Opinnäytetyön tavoite, data warehousen rakentaminen ja käyttöönotto raportoinnissa, onnistui kokonaisuudessaan hyvin ja ilman suurempia ongelmia. Data warehouse -tietokanta saatiin tuotantoon suunnitellun aikataulun puitteissa ja raporttien luominen aloitettua. Kehitys on tietysti jatkuvaa ja erilaista hienosäätöä tehdään koko ajan.

Projektin aikana pääsin tutustumaan muutamaa eri AWS:n palveluun. AWS on valtava sovelluskokonaisuus, jonka kokonainen hallinta lienee yhdelle ihmiselle mahdotonta, mutta pääsin oppimaan palan siitä. Konsultin kanssa työskentely sujui myös hyvin. Meillä oli hyvä kommunikaatio ja ehdin oppia häneltä paljon hyvin lyhyessä ajassa.

Oli mukava päästä työstämään projektia alusta asti. On vaikeampi tulla jatkamaan jonkun muun työtä, koska silloin pitää aina jotenkin yrittää ymmärtää, miksi joku toinen on tehnyt jonkin tietyn valinnan aikaisemmin. Vaikka opinnäytetyö liittyi hyvin tiiviisti muihin työtehtäviini, oli sen kirjallisen osuuden toteuttamiselle vaikea löytää aikaa. Käytännön työn tekemisen jälkeen oli suurempi halu siirtyä eteenpäin uusiin haasteisiin, kuin jäädä kirjoittamaan teoriaa jo tehdystä työstä.

Lähteet

- 1 Heinze J. History of Business Intelligence. 2014; saatavilla: <https://www.better-buys.com/bi/history-of-business-intelligence/>
- 2 Elliott T. Happy Birthday to the “Father of Business Intelligence”. 2013; saatavilla: <https://blogs.sap.com/2013/07/01/happy-birthday-to-the-father-of-business-intelligence/>
- 3 Mertanen M. Business Intelligence kuolee pois. 2017; saatavilla: <https://blog.kauppa-lehti.fi/q-and-a/business-intelligence-kuolee-pois>
- 4 Microsoft - Announcing Power BI general availability coming July 24th. 2015; saatavilla: <https://powerbi.microsoft.com/en-us/blog/announcing-power-bi-general-availability-coming-july-24th/>
- 5 Microsoft - Power BI Pricing. Saatavilla: <https://powerbi.microsoft.com/en-us/pricing/>
- 6 Howson S, Kronz A, Richardson J, Sallam S. Gartner Magic Quadrant for Analytics and Business Intelligence Platforms. 2019.
- 7 Ulag A. Microsoft a Leader in Gartner’s Magic Quadrant for Analytics and BI Platforms for 12 consecutive years. 2019; saatavilla: <https://powerbi.microsoft.com/en-us/blog/microsoft-a-leader-in-gartners-magic-quadrant-for-analytics-and-bi-platforms-for-12-consecutive-years/>
- 8 AWS - Amazon QuickSight Now Generally Available. 2016; saatavilla: <https://aws.amazon.com/about-aws/whats-new/2016/11/amazon-quicksight-now-generally-available/>
- 9 AWS - Amazon QuickSight Pricing. Saatavilla: <https://aws.amazon.com/quicksight/pricing/?nc=sn&loc=4>
- 10 Tableau - 1.0. 2004; saatavilla: <https://www.tableau.com/fast-pace-innovation/1.0>
- 11 Tableau - Pricing for Teams & Organizations. Saatavilla: <https://www.tableau.com/pricing/teams-orgs>

- 12 Kempe S. A Short History of Data Warehousing. 2012, saatavilla: <https://www.dataiversity.net/a-short-history-of-data-warehousing/#>
- 13 Serra J. Why You Need a Data Warehouse. 2013; saatavilla: <http://www.james-serra.com/archive/2013/07/why-you-need-a-data-warehouse/>
- 14 Oracle - Oracle Technology Global Price List. 2007; saatavilla: https://regmedia.co.uk/2008/06/20/oracle_2007_price_list.pdf
- 15 Gartner - Gartner Says Worldwide IaaS Public Cloud Services Market Grew 31.3% in 2018. 2019; saatavilla: <https://www.gartner.com/en/newsroom/press-releases/2019-07-29-gartner-says-worldwide-iaas-public-cloud-services-market-grew-31point3-percent-in-2018>
- 16 Novet J. Amazon's cloud business reports 37% sales growth but misses analysts' estimates. 2019; saatavilla: <https://www.cnbc.com/2019/07/25/aws-earnings-q2-2019.html>
- 17 Bala R, Gill B, Smith D, Wright D. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide. 2019.
- 18 Miller R. Azure revenue continues to slow down for Microsoft. 2019; saatavilla: <https://techcrunch.com/2019/07/18/azure-revenue-continues-to-slow-down-for-microsoft/>
- 19 Barona - Paremman työelämän puolesta. Saatavilla: <https://barona.fi/yrityksille/baronayrityksille/>
- 20 Barona - Etunojassa kohti logistiikan tulevaisuutta. Saatavilla: <https://barona.fi/toimiala/logistiikka/>
- 21 Barr J. Amazon EC2 Beta. 2006; saatavilla: https://aws.amazon.com/blogs/aws/amazon_ec2_beta/
- 22 Barr J. Introducing Amazon RDS – The Amazon Relational Database Service. 2009, saatavilla: <https://aws.amazon.com/blogs/aws/introducing-rds-the-amazon-relational-database-service/>
- 23 Hunt R. Launch – AWS Glue Now Generally Available. 2017; saatavilla: <https://aws.amazon.com/blogs/aws/launch-aws-glue-now-generally-available/>

- 24 Microsoft - Data sources in Power BI Desktop. Saatavilla: <https://docs.microsoft.com/en-us/power-bi/desktop-data-sources>